3D Human Motion Analysis Framework for Shape Similarity and Retrieval

Rim Slama, Hazem Wannous, Mohamed Daoudi

3D Human Motion Analysis Framework for Shape Similarity and Retrieval

Rim Slama[a,b], Hazem Wannous[a,b], Mohamed Daoudi[a,c]

[a] *LIFL laboratory (UMR CNRS 8022), Villeneuve d'Ascq, France*
[b] *University of Lille 1, Villeneuve d'Ascq, France*
[c] *Institut Mines-Télécom / Télécom Lille, Villeneuve d'Ascq, France*

Abstract

3D Shape similarity from video is a challenging problem lying at the heart of many primary research areas in computer graphics and computer vision applications. In this paper, we address within a new framework the problem of 3D shape representation and shape similarity in human video sequences. Our shape representation is formulated using Extremal Human Curve (EHC) descriptor extracted from the body surface. It allows taking benefits from Riemannian geometry in the open curve shape space and therefore computing statistics on it. It also allows subject pose comparison regardless of geometrical transformations and elastic surface change. Shape similarity is performed by an efficient method which takes advantage of a compact EHC representation in open curve shape space and an elastic distance measure. Thanks to these main assets, several important exploitations of the human action analysis are performed: shape similarity computation, video sequence comparison, video segmentation, video clustering, summarization and motion retrieval.

Email addresses: rim.slama@telecom-lille.fr (Rim Slama), hazem.wannous@telecom-lille.fr (Hazem Wannous), mohamed.daoudi@telecom-lille.fr (Mohamed Daoudi)

Experiments on both synthetic and real 3D human video sequences show that our approach provides an accurate static and temporal shape similarity for pose retrieval in video, compared with the state-of-the-art approaches. Moreover, local 3D video retrieval is performed using motion segmentation and dynamic time warping (DTW) algorithm in the feature vector space. The obtained results are promising and show the potential of this approach.
Keywords: Motion analysis, shape similarity, 3D video retrieval, 3D human action.

1. Introduction

While human analysis in 2D image and video has received a great interest during the last two decades, 3D human body is still a little explored field. Relatively few authors have so far reported works on static analysis of 3D human body, but even fewer on 3D human video analysis.

Parallel to this, 3D video sequences of human motion are more and more available. In fact, their acquisition by multiple view reconstruction systems or animation and synthesis approaches [1, 2] received a considerable interest over the past decade, following the pioneering work of Kanade [3].

Most of the recent research topics on 3D video focus mainly on performance, quality improvements and compression methods [4, 2, 5]. Consequently, 3D videos are yet mainly used for display. However, the acquisition of long sequences produces massive amounts of data which necessitates efficient schemes for navigating, browsing, searching, and viewing video data. Hence, we need to develop an efficient and effective descriptor to represent body shape and pose for shape retrieval and video clustering. We also need

a motion retrieval system to look for relevant information quickly.

3D Human body shape similarity is an important area, recently attracted more attention in the field of human-computer interface (HCI) and computer graphics, with many related research studies. Among these, research started with 3D features have been applied for body pose estimation and 3D video analysis.

In this paper, a unified framework providing several processing modules is presented. All viewed within a duality pose/motion approach as summarized in Figure 1 bellow.

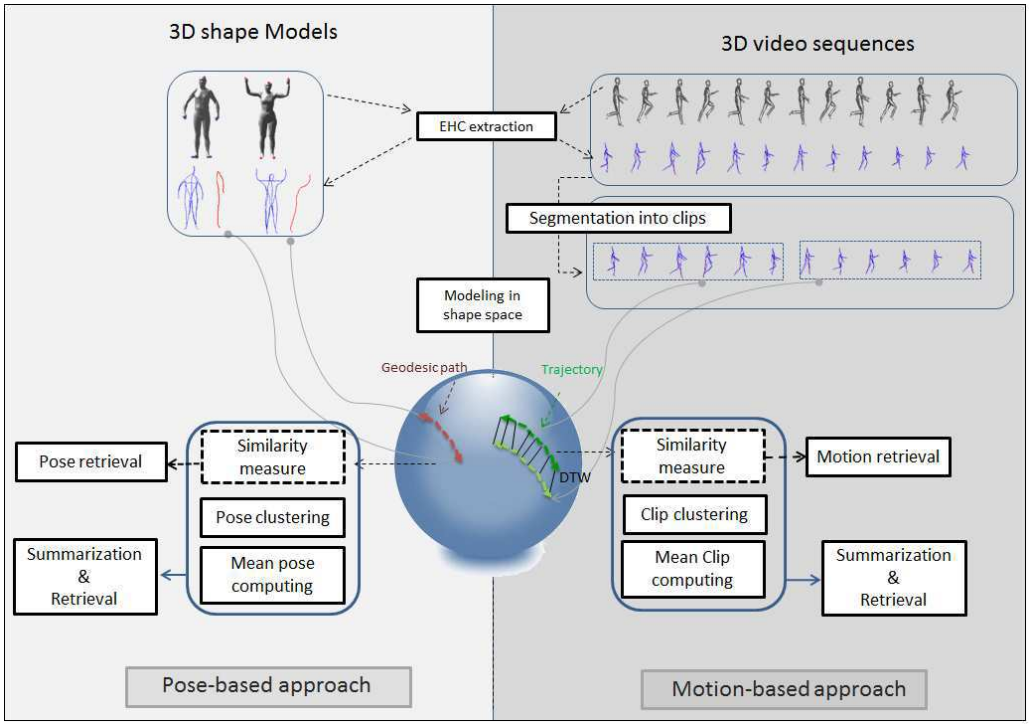


Figure 1: Overview of 3D human motion framework.

We first focus on the analysis of human pose and we propose a novel 3D

human curve-based shape descriptor called Extremal Human Curves (EHC). This descriptor, extracted on body surface, is based on extremal features and geodesics between them. Every 3D mesh is represented by a collection of these open curves. The mesh to mesh comparison is then performed in a Riemannian shape space using an elastic metric between each two correspondent human curves.

At this level, our ultimate goal is to be able to perform reliable reduced representation based on geodesic curves for shape and pose similarity metric. Invariant to pose changes, our EHC descriptor allows pose (and motion) comparison of subjects regardless of translation, rotation and scaling. Such descriptor can be employed not only in pose retrieval for video annotation and concatenation but also in motion retrieval, clustering and activity analysis.

Second, we are interested in the task of video segmentation and comparison between motion segments for video retrieval. As a 3D video of human motion consists of a stream of 3D models, we assume that EHC features are extracted from all 3D shape frames of the sequence, which is further segmented. For direct comparison of video sequences, the motion segmentation can play an important role in the dynamic matching by segmenting automatically the continuous 3D video data into small units describing basic movements, called clips.

For the segmentation of these units, an analysis of minima on motion vector is performed using the metric employed to compare EHC representations. Finally, the motion retrieval is achieved thanks to the dynamic time warping (DTW) algorithm in the feature vector space.

The contributions of this paper are:

- The proposed surface-based shape descriptor called EHC provides a compact representation of the shape. Thereby, reducing both the required space for storage and the time for comparison. As our descriptor is composed of a collection of local human curves, the EHC can find a number of useful applications lying on body part analysis.

- The use of video segmentation allows a semantic analysis of the human motion, within a hierarchical structure of three levels "video-clip-pose".

- The modeling of curves in the shape space manifold allows calculating statistics on shape models and motion clips. Thanks to this latter, templates for the pose/clip are computed as average of a collection of poses/clips. The matching with such templates which represents a class, reduces retrieval complexity algorithm from n to $log(n)$.

- The development of a unified framework, viewed as a duality pose/motion, for several processing modules on video retrieval and understanding, where all use the same features and similarity metric.

The outline of this paper is as follows: Section 2 discusses related works in the area of static and temporal shape similarity and video retrieval. The extremal curve extraction is presented in section 3. Section 4 describes the pose modeling in shape space and the elastic metric used for curve comparison. In section 5, our approach used for motion segmentation and retrieval is presented. Section 6 describes video clustering and summarization for motion understanding. In section 7, evaluation of our framework and experimental results for shape similarity, video segmentation and retrieval are performed. Finally, we conclude by a discussion of the limitations of the approach in

section 8 and a summarizing of our results issues for future works in the conclusion section.

2. Related works

3D shape representation and similarity have been under investigation for a long time in various research fields (computer vision, computer graphics, robotics) and for various applications (3D object recognition, classification, retrieval). We address below, the most relevant works related to our approach, which only utilize the full-reconstructed 3D data for shape similarity in 3D human video.

Most works which address this problem evaluate a similarity metric on static shape descriptors based on the surface or on the volume. Others propose to extend the static approaches to temporal shape descriptors.

2.1. Static descriptors

Some of widely used 3D object representation approaches include: spin images, spherical harmonics, shape context and shape distribution. Johnson et al. [6] propose spin image descriptor, encoding the density of mesh vertices into 2D histogram. Osada et al. [7] use a Shape Distribution, by computing the distance between random points on the surface. Ankerst et al. [8] represent the shape as a volume sampling spherical histogram by partitioning the space containing an object into disjoint cells corresponding to the bins of the histogram. This later is extended with color information by Huang et al. [9]. A similar representation to the Shape Histogram is presented by Kortgen et al. [10] as 3D extended shape context. Kazhdan et al. [11] apply spherical harmonics to describe an object by a set of spherical basis functions

representing the shape histogram in a rotation-invariant manner. These approaches use global features to characterize the overall shape and provide a coarse description, that is insufficient to distinguish similarity in 3D video sequence of an object having the same global properties in the time. A comparison of these shape descriptors combined with self-similarities is made by Huang et al. [12].

Other works on the 3D shape similarity can be found in the literature, where surface-based descriptors are often used with a step of features detection. The advantage of these features is that their detection is invariant to pose change. The extremities can be considered as the one among the most important features for the 3D objects. They can be used for extracting a topology description of the object like Reeb-graph descriptor [13] or closed surface-based curves [14, 15, 16]. The extraction and the matching of these features have been widely investigated using different scalar functions from geodesic distances to heat-kernel [17, 18, 19]. Tabia et al. [14] propose to extract arbitrarily closed curves amounting from feature points and use a geodesic distance between curves for 3D object classification. Elkhoury et al. [15] extract the same closed curves but they use heat-kernel distance in the 3D object retrieval process.

2.2. Temporal descriptors

Since significant progress in multiple view reconstruction techniques has been made, 3D video sequences of human motion are more and more available. However, the need for handling and processing such data led to several approaches using temporal shape representation and matching.

Huang et al. [12] extend the use of static descriptors to temporal ones for

frame retrieval, in a 3D human video, using time filtering and shape flows obtained via invariant-rotation shape histograms. Such approaches give a good shape descriptor but usually do not capture any geometrical information about the 3D human body pose and joint positions/orientations. This prevents using them in certain applications that require accurate estimation of the pose (and the joints in some cases) of the body parts. The temporal similarity in 3D video is addressed also in the case of skeletal motion and is evaluated from difference in joint angle or position together with velocity and acceleration [20]. Huang et al. [21] demonstrate that skeleton-based Reeb-Graph descriptor has a good performance in the task of finding similar poses of the same person in 3D video. Shape similarity is also used for solving the problem of video retrieval by matching frames and comparing correspondent ones using a specified metric. In Yamasaki et al., [22] the modified shape distribution histogram is employed as feature representation of 3D models. The similar motion retrieval is realized by Dynamic Programming matching using the feature vectors and Euclidean distance. The Dynamic Time Warping algorithm (DTW), based on Dynamic Programming and some restrictions, was also widely used to resolve the problem of temporal alignment. Given two time series with different size, DTW finds an optimal match measuring the similarity between these sequences which may vary in time or speed. Thereby, by a frame descriptor and the temporal alignment using DTW, many authors succeed to perform action recognition or sequence matching for indexing [23, 24, 25].

Recently, Tung et al. [13] propose a topology dictionary for video understanding and summarizing. Using the Multi-resolution Reeb Graph as

a relevant descriptor for the shape in video stream for clustering. In this approach, they perform a clustering of the video frames into pose clusters and then they represent the whole sequence with a Markov motion graph in order to model the topology change states.

From the above review, we can identify certain issues in order to consider in our approach. Most of these works have attempted to use global description of the model ignoring the local details. The similarity metric is usually calculated directly on descriptors whereas the notion of motion is incorporated by time convolution of the distance metric itself computed from static poses. The video sequence is considered as a succession of frames in time and not a succession of elementary motions (or gestures).

On one hand, the extremities feature points used in many state-of-the-art algorithms can be considered as an important compact semantic representation of human posture. On the other hand, the shape analysis of curves extracted from human body mesh allows representing the shape variations. Choosing some representative curves of the body surface may provide an efficient and a compact representation of human shape.

Our approach has several benefits: (1) the EHC descriptor can be considered as a surface skeletal based representation, which allows to describe surface deformations of the human posture. As it is composed by a collection of local extremal open 3D curves, a body part representation can be performed; (2) the motion analysis is incorporated in two ways, firstly by time convolution of the distance metric vectors for pose retrieval in video sequence, and secondly by employing motion segmentation and the notion of

clips; (3) the video segmentation allows the localization of transition states in the video, in order to analyze the local dynamic of the motion, representing an atomic action or gesture; and (4) an original idea is proposed to represent a clip as a trajectory composed of a collection of successive frames viewed as points in shape space. Finally, the video segmentation and clustering are exploited in content-based summarization and motion retrieval.

3. Extremal Curves

We aim to represent a body shape as a skeleton based shape representation. This skeleton will be extracted on the surface of the mesh by connecting features located on the extremities of the body. The main idea behind the use of this representation is to analyze pose variation with elastic deformation of the body, using representative curves on the surface.

3.1. Feature point detection

Feature points refer to the points of the surface located at the extremity of its prominent components. They are useful in many applications, including deformation transfer, mesh retrieval, texture mapping and segmentation. In our approach, feature points are used to represent a new pose descriptor based on curves connecting each two extremities. Several approaches have been proposed in the literature to extract feature points; Mortara et al. [26] select as features points the vertices where Gaussian curvature exceeds a given threshold. Unfortunately, this method can miss feature points because of the threshold parameter and cannot resolve extraction on constant curvature areas. Katz et al. [27] develop an algorithm based on multidimensional scaling, in quadratic execution complexity. Another approach more robust,

is proposed by Tierny et al. [28] to detect extremal points, based on geodesic distance evaluation. This approach is used successfully to detect the body extremities, since it is stable and invariant to geometrical transformations and model pose. The extraction process can be summarized as the following:

Let v_1 and v_2 be the most geodesic distant vertices on a connected triangulated surface S of a human body. These two vertices are the farthest on S, and can be computed using Tree Diameter algorithm (Lazarus et al. [29]). Now, let f_1 and f_2 be two scalar functions defined on each vertex v of the surface S as follows:

$$f_1(v) = g(v, v_1) \setminus f_2(v) = g(v, v_2) \tag{1}$$

where $g(x, y)$ is the geodesic distance between points x and y on the surface. Let E_1 and E_2 be respectively the sets of extrema vertices (minima and maxima) of f_1 and f_2 on S (calculated in a predefined neighbourhood). We define the set of feature points of the surface of human body S as the intersection of E_1 and E_2. Concretely, we perform a crossed analysis in order to purge non-isolated extrema, as illustrated in Figure 2 (top). The f_1 local extrema are displayed in blue color, f_2 local extrema are displayed in red color and feature points resulting from their intersection are displayed in mallow color. Figure 2 (bottom) shows different persons from three different datasets where feature extraction is stable despite change in shape, pose and clothing for each actor.

3.2. Body curve extraction

Let M be a body surface and $E = \{e_1, e_2, e_3, e_4, e_5\}$ a set of feature points on the body representing the output of the extraction process. Let β denotes

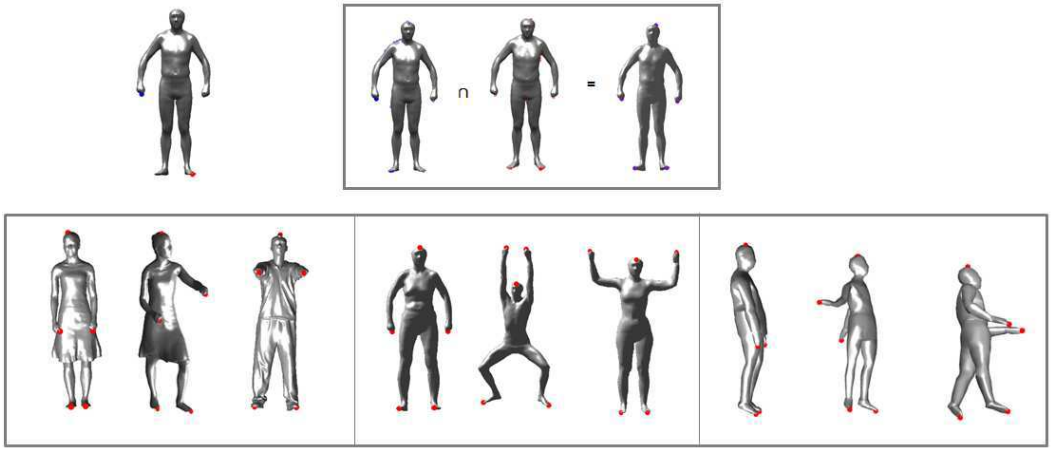


Figure 2: Extremity points of the 3D human body. (top) extracting process, (bottom) different human body subjects in different poses.

the open curve on M which joints two feature points of M $\{e_i, e_j\}$. To obtain β, we seek for geodesic path P_{ij}, whose length is shortest while passing through the surface of the mesh, between e_i and e_j. We repeat this step to extract extremal curves from the body surface ten times so that we do all possible paths between elements of E. As illustrated in the top of Figure 3, the body posture is approximated by using these extremal curves $M \sim \bigcup \beta_{ij}$, and we can categorize these curves into 5 categories (Figure 3 bottom):

- Curves connecting hand and foot on the same side: for controlling the movement of the left/right half of the body.
- Curves between hands and between feet: for controlling the movement of the upper/lower body.
- Curves connecting crossed hand and foot: for controlling the movement

of the crossed limbs.

- Curves between head and feet: for controlling the movement of right/left foot.

- Curves between head and hands: for controlling the movement of right/left hands.

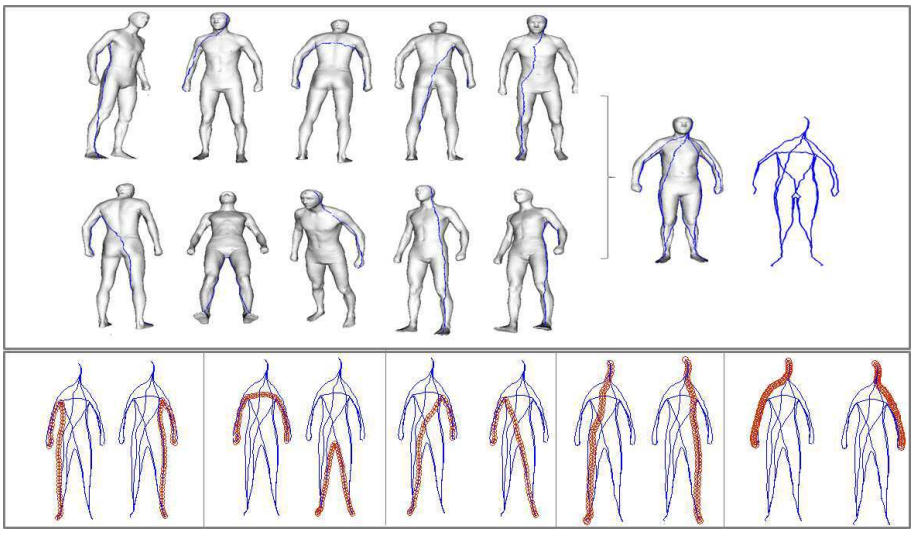


Figure 3: Body representation as a collection of extremal curves.

Note that modeling objects with curves is recently carried out for several applications; Abdelkader et al. [30] use closed curves extracted from human silhouettes to characterize human poses in 2D videos for action recognition. Drira et al. [31] use open curves extracted from nose tip and face surface as a surface parametrization for 3D face recognition.

In our approach, we have chosen to represent the body pose by a collection of curves for two reasons. Firstly, these curves connect limbs and

give obviously a good representation of the body shape and pose, using a reduced representation of the mesh surface. Secondly, this representation allows studying the shape variation using Riemannian geometry by projecting these curves in the shape space of curves and using its elastic metric introduced by Joshi et al. [32].

4. Pose modeling in shape space

In order to compare the similarity between two human body postures, we must quantify the change of shape between correspondent curves. To do this, the metric used to compare shape of curves can be computed inside an open curve shape space.

In the last few years, many approaches have been developed to analyze shapes of 2-D curves. We can cite approaches based on Fourier descriptors, moments or the median axis. More recent works in this area consider a formal definition of shape spaces as a Riemannian manifold of infinite dimension on which they can use the classic tools for statistical analysis. The recent results of Michor et al. [33], Klassen et al. [34] and Yezzi et al. [35] show the efficiency of this approach for 2-D curves. Joshi et al. [32] have recently proposed a generalization of this work to the case of curves defined in $\mathbb{R}^n$. We adopt this work to our problem since our 3-D curves are defined in $\mathbb{R}^3$.

4.1. Elastic distance

While human body is an elastic shape, its surface can be simply affected by a stretch (raising hand) or a bind (squatting). In order to analyze human curves independently to this elasticity, an elastic metric is needed within a shape space framework.

Let $\beta : I \rightarrow \mathbb{R}^3$, for $I = [0, 1]$, represents an extremal curve obtained as described above. To analyze its shape, we shall represent it mathematically using a *square-root velocity function* (SRVF), denoted by $q(t) \doteq \dot{\beta}(t)/\sqrt{\|\dot{\beta}(t)\|}$. $q(t)$ is a special function introduced by Joshi et al.[32] that captures the shape of β and is particularly convenient for shape analysis.

The set of all unit-length curves in $\mathbb{R}^3$ is given by $\mathcal{C} = \{q : I \rightarrow \mathbb{R}^3 | \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^3)$, where using $\mathbb{L}^2$-metric on its tangent spaces, $\mathcal{C}$ becomes a Riemannian manifold.

Proposition 4.1. *Having two open curves represented by their SRVF, q_1 and q_2, the shortest geodesic between them in the shape space of open curves is given by:* $\alpha(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2^*\right)$,
and the geodesic distance is given by: $d_s(q_1, q_2) \doteq cos^{-1}(\langle q_1, q_2^* \rangle)$.

where q_2^* is the optimal element associated with the optimal rotation O^* and re-parametrization γ^* of the second curve.

This defined distance allows comparing shape curves regardless of isometric and elastic deformation. In Figure 4, a geodesic path between each corresponding two extremal curves, taken from two human bodies doing different poses, is computed in shape space.

For the left model, the person's arm is down and for the right model it is raised. The geodesic path between each two curves is shown in the shape space. This evolution looks very natural under the elastic matching.

4.2. Static shape similarity

The elastic metric applied on extremal curve-based descriptors can be used to define a similarity measure. Given two 3D meshes x, y and their

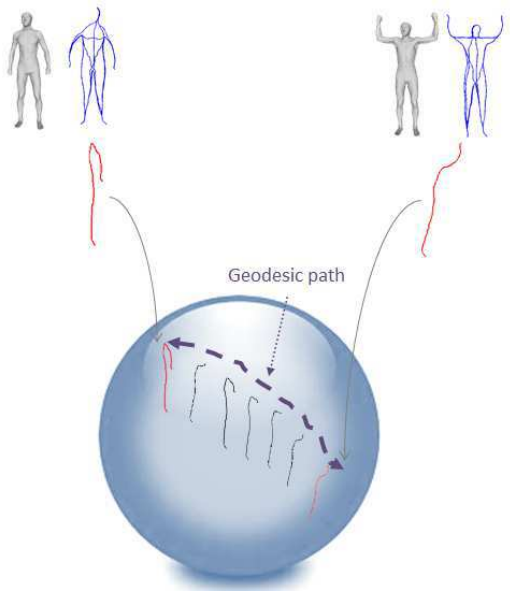


Figure 4: Geodesic path between two extremal human curves of neutral pose with raised hands.

descriptors $x' = \{q_1^x, q_2^x, q_3^x, ..., q_N^x\}$ and $y' = \{q_1^y, q_2^y, q_3^y, ..., q_N^y\}$, the mesh-to-mesh similarity can be represented by the curve pairwise distances and can be defined as follows:

$$s(x, y) = d(x', y') \ , \tag{2}$$

$$d(x', y') = \frac{\sum_{i=1}^{N} d(\beta_i^x, \beta_i^y)}{N} = \frac{\sum_{i=1}^{N} d_s(q_i^x, q_i^y)}{N}. \tag{3}$$

where N is the number of curves used to describe the mesh. The mean of curve distances between two descriptors captures the similarity between their mesh poses. In case of shape change in even one curve, the global distance is affected and it increases indicating that the poses are different. In order to have a global distance, an arithmetic distance can be computed in order to compare human poses.

4.3. Average poses

The use of EHC descriptor to represent the human pose by a collection of 3D open curves allows analyzing the human shape using the geometrical framework. It also allows computing some related statistics like "average" of several extremal human curves. Such an average, called Karcher Mean, is introduced by Srivastava et al. [36]. It can be computed between different poses to represent the intermediate pose, or between similar poses done by several actors to represent a template for similar poses.

We are interested in defining a notion of "mean" for a given set of human postures in the same cluster of poses for the goal of fast pose retrieval.

To compute the average of EHC representation, we need only to know how to compute an average for one extremal human curve. The Riemannian structure defined on the shape space $\mathcal{S}$ enables us to perform such statistical analysis for computing average and variance for each 3D open curve on body surface. The intrinsic average or the Karcher Mean utilizes the intrinsic geometry of the manifold to define and compute a mean on that manifold. In order to calculate the Karcher Mean of extremal human curves $\{q_1^\alpha, ..., q_n^\alpha\}$ in $\mathcal{S}$, we define the variance function as:

$$\mathcal{V} : \mathcal{S} \to \mathbb{R}, \mathcal{V}(\mu) = \sum_{i=1}^{n} d_{\mathcal{S}}(q_i^\alpha, q_j^\alpha)^2 \tag{4}$$

The Karcher Mean is then defined by:

$$\overline{q^\alpha} = \arg\min_{\mu \in \mathcal{S}} \mathcal{V}(\mu) \tag{5}$$

The gradient of $\mathcal{V}$ is used in the tangent space $T_\mu(\mathcal{S})$ to iteratively update the current mean μ. $\overline{q^\alpha}$ is an element of $\mathcal{S}$ that has the smallest geodesic

path length from all given extremal human curves for the index α.

An example of using the Karcher Mean to compute average curve for 6 extremal human curves connecting hand and foot from the same side is shown in the top of Figure 5, and several examples of using the Karcher mean to compute average EHC representation are shown in the bottom of this figure.

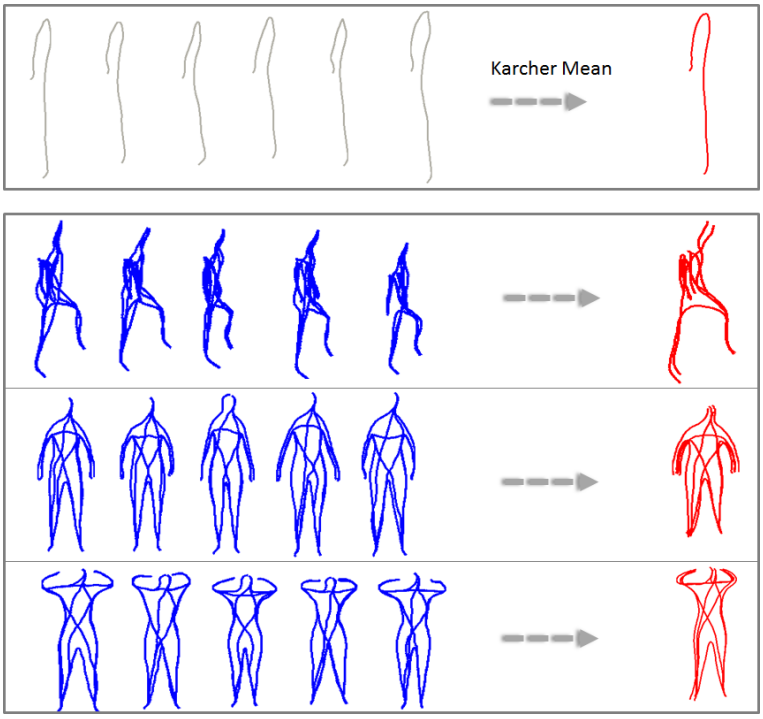


Figure 5: Example of Karcher Mean computation. (top) Mean curve for six extremal human curves: curve connecting hand and foot from the same side. (bottom) Example of average poses computed using Karcher mean.

5. Motion segmentation and matching

Based on our EHC representation of the shape model, it is possible to compare two video sequences by matching all pairwise correspondent extremal curves inside their frames, using the geodesic distance in the shape space. However, a sequence of human action can be composed of several distinct actions, and each one can be repeated several times. Therefore, the motion segmentation can play an important role in the dynamic matching by dividing the whole 3D video data into small, meaningful and manageable elementary actions called clips. EHC descriptor will be employed to segment continuous sequences into clips.

5.1. Motion segmentation

Video segmentation has been studied for various applications, such as gesture recognition, motion synthesis and indexing, browsing and retrieval. A vast amount of works in video segmentation has been performed for 2D video [37], where usually the object segmentation is firstly performed before the movement analysis. In Rui et al. [38], an optical flow of moving objects is used and motion discontinuities in trajectories of basis coefficient over time are detected. However, in Wang et al. [39], break points were considered as local minima in motion and local maxima in direction change.

Motion segmentation is strongly applied in several algorithms using 3D motion capture feature points trackable within the whole sequence, to segment the video. Detected local minima in motion (Shiratori et al. [40]) or extrema (Kahol et al. [41]) are used in motion segmentation for kinematic parameters.

Most of works on the 3D video segmentation use the motion capture data, and very few of them were applied to dynamic 3D mesh. One of them is presented by Xu et al. [42], where a histogram of distance among vertexes on 3D mesh is generated to perform the segmentation through thresholding step defined empirically. In Yamasaki et al. [43], the motion segmentation is automatically conducted by analyzing the degree of motion using modified shape distribution for mainly japanese dances. These sequences of motion are paused for a moment and then they are consider as segmentation points. Huang et al. [44] propose an automatic key-frame extraction method for 3D video summarization. To do so, they compute the self similarity matrix using volume-sampling spherical Shape Histogram descriptor. Then, they construct a graph based on this self similarity matrix and define a set of key frames as the shortest path of this graph.

In our work, we propose an approach fully automatic to segment a 3D video efficiently without making neither thresholding step nor assumption on the motion's nature. In motion segmentation, the purpose is to split automatically the continuous sequence into segments which exhibit basic movements, called clips. As we need to extract meaningful clips, the segmentation is overly fine and can be considered as finding the alphabet of motion. For a meaningful segmentation, motion speed is an important factor. In fact, when human changes motion type or direction, the motion speed becomes small and this results in dips in velocity. We exploit this latter by finding the local minima for the change in type of motion and local maxima for the change in direction. The extrema detected on velocity curve should be selected as segment points (see Figure 6). We show frames detected as maxima (the ac-

tor changes the foot's direction) on the top of the plot, and frames detected as minima (the actor raise the other foot) on the bottom. In this work, we consider only the change in type of motion as a meaningful clip. Thus, clips with slight variations and a small number of frames are avoided.

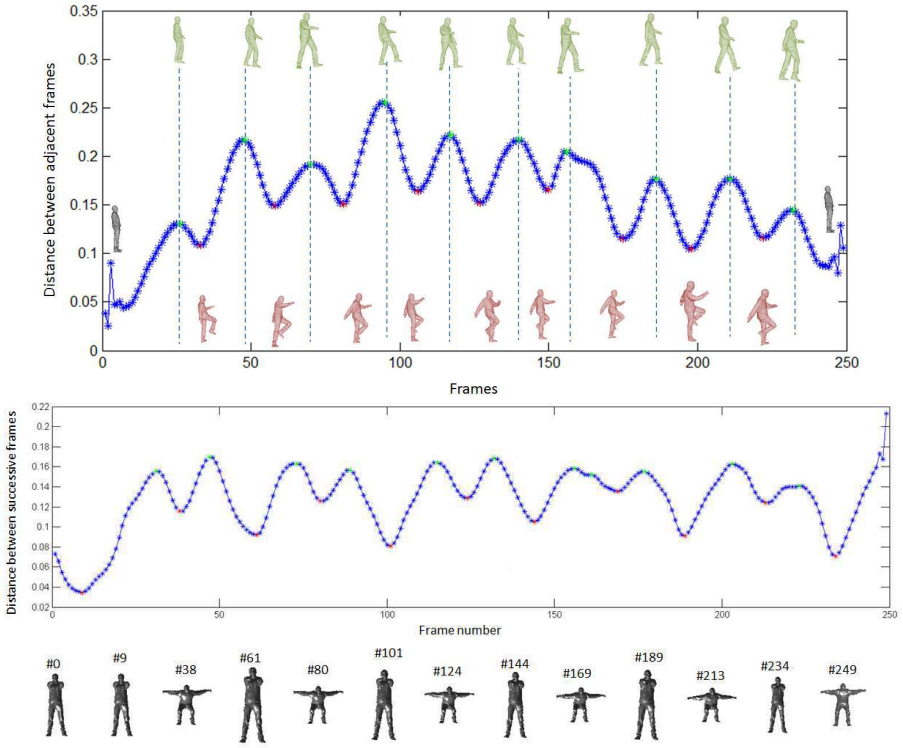


Figure 6: Segmentation of a 3D sequence into motion clips. Feature vector and detected frames as local extrema are presented at the top of the figure and detected frames as minima are at the bottom.

Note that optimum local minimum, that detect precise break points where the motion changes, is selected in a predefined neighbourhood. For this raison, we fix a size of window to test the efficiency of the local minimum in

this condition. To calculate the speed variation, distance between each two successive EHC in the sequence is computed. The variations of the sequence are represented in vector of speed and a further smoothing filter is applied to obtain the final degree of motion vector.

5.2. Clip matching

To seek for similar clips, we need to encode gestures in a specific representation that we can compare regardless to certain variations. In fact, two motions are considered similar even if there are changes in the shape of the actor and the speed of the action execution. This problem is similar to time-series retrieval where a distance metric is used to look for, in a database, the sequences whose distance to the query is below a threshold value. Each clip is represented as a temporal sequence of human poses, characterized by EHC representation associated to shape model. Then, extremal curves are tracked in each sequence to characterize a trajectory of each curve in the shape space as illustrated in Figure 7 (top). Finally, the trajectories of each curve are matched and a similarity score is obtained. However, due to the variations in execution rates of the same clip, two trajectories do not necessarily have the same length. Therefore, a temporal alignment of these trajectories is crucial before computing the global similarity measure, as shown in Figure 7 (bottom).

In order to solve the temporal variation problem, we use DTW algorithm (Giorgino et al. [45]). This algorithm is used to find optimal non-linear warping function to match a given time-series with another one, while adhering to certain restrictions such as the monotonicity of the warping in the time domain. The optimization process is usually performed using dynamic

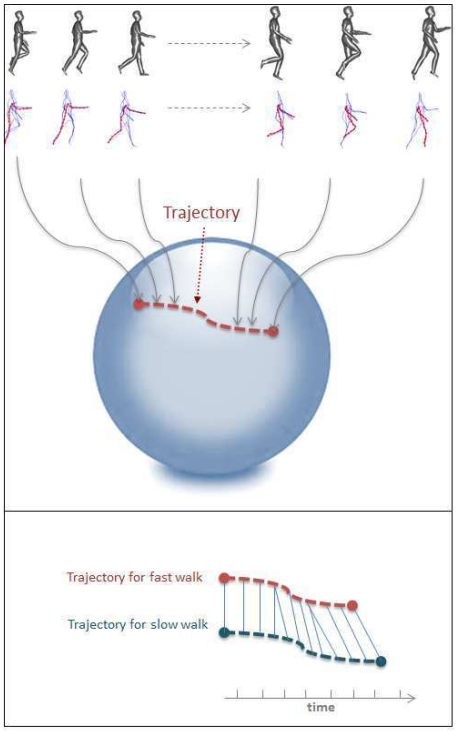


Figure 7: Graphical illustration of a sequence of shapes obtained during a walking action (top). Alignment process between trajectories of same curve index using DTW (bottom).

programming approaches given a measure of similarity between the features of the two sequences at different time instants. The global accumulated costs along the path define a global distance between the query clip and the motion segments found in the database. Since DTW can operate with any measure of similarity between different temporal features, we adapt it to features that reside on Riemannian manifolds. Hence, we use the geodesic distance between different shape points $d_s(q_i, q_j)$ as a distance function between the shape features at different time instants.

In practice, the first step is to follow independently curve variation in time resulting on N trajectories in the shape space. In fact, each frame in the 3D video sequence can be represented by a predetermined number (N) of extremal curves, splitting the sequence into N parts, where each one represents the trajectory of an open curve in the shape space. Then, DTW will be applied in the feature space for each tracked curve index. The distance between two clips is then the average distance given by each comparison between corresponding trajectories.

5.3. Average clip

Based on the two algorithms, Karcher Mean and DTW, we can extend the notion of "mean" of a set of human poses to the "mean" of trajectories of poses in order to compute an "average" of several clips.

Let N be the number of clips represented by N trajectories $T_1, T_2 \cdots T_N$. For a specific human curve index, we look for the mean trajectory that has the minimum distance to the all N trajectories.As shown in Algorithm 1, the mean trajectory is given by computing the non-linear warping functions and setting iteratively the template as the Karcher Mean of the N warped trajectories represented in the Riemannian shape space.

6. Video summarization and retrieval

In order to represent compactly a video sequence, we need to know how to exploit the redundancy of information over time. However, when this information should be extracted from motion and not from frames separately, the challenge is then about complex matching processes required to find geometric relations between consecutive data stream elements. We therefore

Algorithm 1 Computing trajectory template

Require: N trajectories from N clips $T_1, T_2 \cdots T_N$

Initialization: chose randomly one of the N input trajectories as an initial guess of the mean trajectory T_{mean}

repeat

for i=1 : N **do**

find optimal path p^* using DTW to warp T_i to T_{mean}

end for

Update T_{mean} as the Karcher Mean of all N warped trajectories

until Convergence

propose to use EHC to represent a pose and a trajectory as key descriptors characterizing geometric data stream. Based on EHC representation, we develop several processing modules as clustering, summarization and retrieval.

6.1. Data clustering

Let V denotes a video stream of human sequence containing elements $\{e_i\}_{i=1...k}$, where e can be a frame or a clip. To cluster V, the data set is recursively split into subsets C_t and R_t as described in the following recursive algorithm:

Algorithm 2 Data clustering

Require: $V\{e_i\}_{i=1...k}$;

Ensure: $C_0 = \varnothing$; $R_0 = \{e_1, \ldots, e_k\}$;

if $(R_t \neq \varnothing)\&\&(t <= k)$ **then**

$C_t = \{f \in R_{t-1} : dist(e_t, f) < Th\}$;

$R_t = R_{t-1} \setminus C_t$;

end if

The result of clustering is contained in $C_{t=1..k}$ where C_t is a subset of V

representing a cluster containing similar elements to e_t. For each iteration of clustering steps, from $t = 1 \cdots K$, the closest matches to e_t are retrieved and indexed with the same cluster reference as e_t. Any visited element e_t already assigned to a cluster in C during iteration step is considered as already classified and is not processed subsequently. We regroup not empty sub sets C_t in l clusters $\{c_1, ..., c_l\}$ (with $l \leqslant k$). Similarities between elements of V are evaluated using a similarity distance $dist$ allowing to compare the elements of V. The threshold Th is defined experimentally .

If we consider the video V as a long stream of 3D meshes, the clusters that should be obtained must gather models with similar poses. In this case, the EHC feature vector is used as an abstraction for every mesh and the similarity distance is the elastic metric computed between each pair of human poses. Motion can be incorporated in this similarity by applying a simple time filter on static similarity measure with a window size chosen experimentally [46]. The use of temporal filter integrates consecutive frames in a fixed time window, thus allowing the detection of individual poses while taking into account smooth transitions. Note that the pose invariance property of the EHC allows us to compare poses (and motions) of subjects regardless of translation, rotation and scaling .

The video V can also be considered as a stream of clips resulting from the video segmentation approach and clusters here gathers clips with similar repeated atomic actions. In this case: (1) the feature vector used as abstraction for each clip is a trajectory on shape space of extremal human curves; and (2) the similarity distance, used to compare clips, is based on the DTW algorithm.

6.2. Content-based summarization

Our approach for video summarization is based on three steps: First, the whole video is segmented and clustered into several clusters of clips. Second, only the most significant clip (the nearest one to all cluster elements) of each cluster is kept. Third, we construct a subsequence, from the starting video, where this representative clips of each cluster are concatenated. Finally, This new subsequence, is clustered into clusters of poses, and only most representative poses are kept to describe the dataset.

This summarization allows a reduction of dimension for the original dataset where we can display only main clips if we stop on third step, or to display key frames if we continue summarization process until pose clustering.

6.3. Pose and motion Retrieval

As in a classical retrieval procedure, in response to a given query, an ordered list of responses that the algorithm found nearest to the query is given. Then to evaluate the algorithm, this ranked list is analyzed. Whatever the given query, pose or clip, the crucial point in the retrieval system is the notion of "similarity" employed to compare different objects.

For content-based pose retrieval, thanks to the static shape similarity, we are able to compare human poses using their extremal human curve descriptors and decide if two poses are similar or not. In this scenario, the query consists of a 3D human shape model in a given pose and the response is 3D human bodies more similar in pose to the query. We advocate the usage of the EHC to represent the 3D human shape model in a given pose and then comparison between each pair of models using the elastic metric defined in

the Proposition 4.1. This system can find a number of utilities like pose-based searching and facilitate retrieval of efficient information as subjects in same poses in the database of 3D models scanned in different poses [47, 48].

Note also that identifying frames with similar shape and pose can be used potentially for concatenative human motion synthesis. Concatenate existing 3D video sequences allows the construction of a novel character animation. A good descriptor that much correctly correspondent frames allows the synthesis of videos with smooth transitions and finding best frames to summarize the video. However, extension of static shape descriptor to include temporal motion information is required to remove the ambiguities inherent in static shape descriptor for comparing 3D video sequences of similar shape. Therefore, the static shape descriptor can be extended to the time domain by applying a simple time filter with a window size like $2N_t + 1$. This time filter is a way of incorporating motion in the similarity measure, as so-called temporal similarity, also used by Huang et al. [12]. The temporal similarity is presented in the following Equation:

$$s_{ij}^{t} = \frac{1}{2N_t + 1} \sum_{k=-N_t}^{N_t} s(i + k, j + j) \tag{6}$$

where s is the frame-to-frame similarity matrix and N_t is a time filter with window size $2N_t + 1$.

For content-based motion retrieval, we advocate the usage of the EHC representation, where a query consists of a trajectories representing a clip on the shape space. As response to this specific query, our approach looks in the sequence for most similar trajectories and returns an ordered list of similar ones using the process of motion clip explained in section 5.2.

7. Experimental results

To show the practical relevance of our method, we perform an experimental evaluation on several databases (summarized in Table 1) and compare it to the most efficient descriptors of the state-of-the-art methods. We first evaluate our descriptor for shape similarity application over public static shape database [48] and evaluate the results against Spherical Harmonic descriptor [11]. Secondly, we measure the efficiency of our descriptor to capture the shape similarity in 3D video sequences of different actors and motions from other public 3D synthetic [12] and real [49, 50] video databases. We evaluate this later against Temporal Shape Histogram [12], Multi-resolution Reeb-graph [21] and other classic shape descriptors, using provided Ground Truth. Motion segmentation into clips and clip matching performance are tested on several video sequences of different people doing different motions. Finally, we evaluate our clustering and summarization approach for pose/clip-based video retrieval.

7.1. Feature matching

The extraction and comparison of our curves requires the identification of feature end-points as head, right/left hand and right/left foot, which is not affordable in practice. This requirement is important to perform the curve matching separately between models. In order to overcome this problem, our method is based on two benefits from the morphology of the human body. First, we deduce that geodesic path connecting each one of the hand end-points and the head end-point is shortest among all possible geodesics between the five end-points. Second, the geodesic path connecting right hand

Dataset	Motions/Poses	Number of frames
Dataset (1) [48]: 144 subjects (59 men/55 women)	18 static poses (1 neutral done by all subjects and 17 other different poses)	⊘
Dataset (2) [12]: 14 people (10 men and 4 women)	28 motions: sneak, walk (slow, fast, turn left/right, circle left/right, cool, cowboy, elderly, tired, macho, march, mickey, sexy,dainty), run (slow, fast,turn right/left, circle left/right), sprint, vogue, faint, rockn'roll, shoot.	392 seq, 39200f (100f per seq.)
Dataset (3) [49]: 3 people (2 men and 1 woman)	6 motions: 2×cran, 2×marche, 2×squat, 1×handstand, 1×samba, 1×swing.	1582f (on average 226 ± 48 per seq.)
Dataset (4) [50]: Roxanne	Game character motion: walk	32 f

Table 1: Summarization of data used for all experimental tests.

to left foot end-points or left hand to right foot end-points is the longest. The first observation allows to identify precisely the end-point corresponding to the head, the two end-points connected to this later corresponding to the hands without distinguishing between right and left. The second one allows the identification of the couple of hand/foot as corresponding to same side of the body without distinguishing between right and left. A prior knowledge on the direction of the posture of the human body for static pose and in the starting frame for video sequence has allowed to distinguish between left and

right. Once the end-points are correctly detected from the starting frame in the video sequence, a simple algorithm of end-point tracking over time is performed.

7.2. Static shape similarity

The protocol and the dataset used to validate the experiments are firstly presented and then, the results following this protocol are analyzed and compared to those obtained by other approaches.

7.2.1. Evaluation methodology

To assess the performance of the EHC for static shape similarity, several experiments were performed on a statistical shape database [48]. This database, summarized in Table 1 (1^{st} row), is challenging for human body shape and pose retrieval as it is realistic shape database captured with a 3D laser scanner. It contains more than hundred subjects doing more than thirty different poses. We perform our descriptor on a subset of 338 shape models obtained from 144 subjects 59 male and 55 female aged between 17 and 61 years. There are 18 consistent poses (p0, p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, p11, p12, p13, p16, p28, p29, p32). Some poses are illustrated in Figure 8. Each pose represents a class where at least 4 different subjects do the same pose.

For evaluation, we use Recall/Precision plot in addition to the three statistics which indicate the percentage of the top K matches that belong to the same pose class as the query pose:

- The nearest neighbor statistic (NN): it provides an indication to how well a nearest neighbor classifier would perform (here $K = 1$).

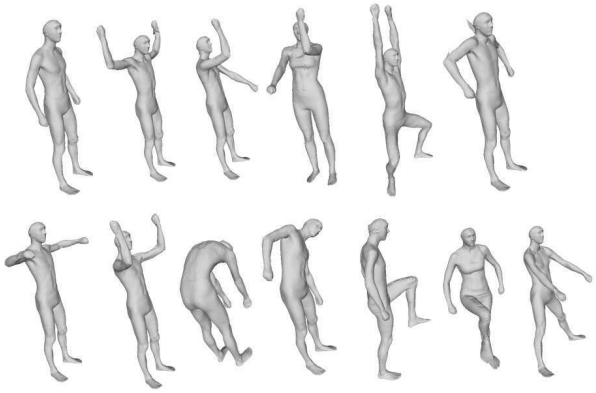


Figure 8: Example of body poses in the static human dataset [48].

- The first tier statistic (FT): it indicates the recall for the smallest K that could possibly include 100% of the models in the query class.
- The second tier statistic (ST): it provides the same type of result, but it is a little less stringent (i.e., K is twice as big).
- E-Measures: it is a composite measure of precision and recall for a fixed number of retrieved results.

We note here that these statistics will be used for static and video retrieval evaluations.

7.2.2. Curve selection

From five feature endpoints, we have extracted ten extremal curves representing the human body shape model. According to the human poses, extremal curves exhibit different performance and some curves are more efficient to capture the shape similarity between two poses. Our shape descriptor

can be seen as a concatenation of ten curve representations and the similarity between two shape models doing two different poses, is represented by a vector of ten elastic distance values. Before all tests, we analyze the performance of all possible combinations of curves on the shape similarity measurements. A Sequential Forward Selection method, applied on elastic distance values and coupled with ST statistic, has been used to select the best combination of curves among all possible ones (1013 combinations according to Eq. 7):

$$\sum_{k=2}^{n} C_n^k = \sum_{k=2}^{n} \frac{n!}{k!(n-k)!} \tag{7}$$

where n is equal to 10 and it represent the number of curves.
Experiment of pose-based retrieval on the dataset [48] shows that the best combination is obtained by the five curves: right hand to right foot, left hand to left foot, left hand to right hand, left foot to right foot, and head to the right foot (Figure 9).

The selected five curves seem to be the most stable ones and they are sufficient to represent at best the body like a skeleton on the surface. Therefore, the elimination of five curves allows to eliminate the ambiguity due to the redundancy of some curves on the body parts.

7.2.3. Result analysis

The self similarity matrix obtained from the mean elastic distance of the five selected curves is shown in the Figure 10.

This matrix demonstrates that similar poses have a small distance (cold color) and that this distance increases with the degree of the change between poses (hot color). This allows pose classification or pose retrieval by

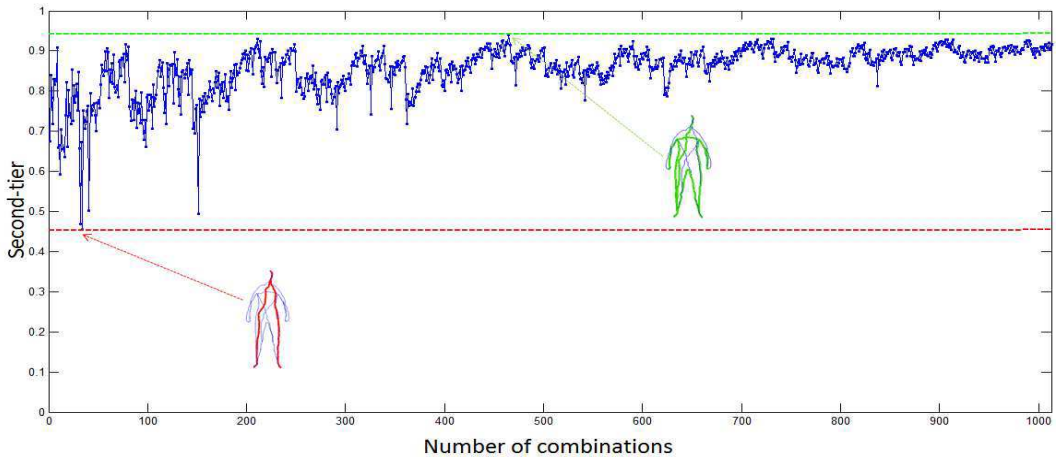


Figure 9: Second-Tier statistic for all combinations of curves. The best combination is obtained by 5 curves (green) and the worst combination is obtained by 2 curves (red).

comparing models using their extremal curve representation and the elastic metric.

From a quantitative point of view, we present the Recall/Precision plot obtained by EHC compared to the popular Spherical Harmonic (SH) descriptor with optimal parameter setting ($N_s = 32$ and $N_b = 16$) [8]. This plot and accuracy rates (NN, FT and ST) reported in Table 2 show that our approach provides better retrieval precision. EHC using only the five selected curves outperforms SH and EHC using the 10 curves to retrieve models with the same pose.

Note finally that the accuracies of retrieval ranks for some poses are relatively low. Such ambiguities can be noticed in the case of comparison between neutral pose and a pose where subjects just twist their body to the left, or

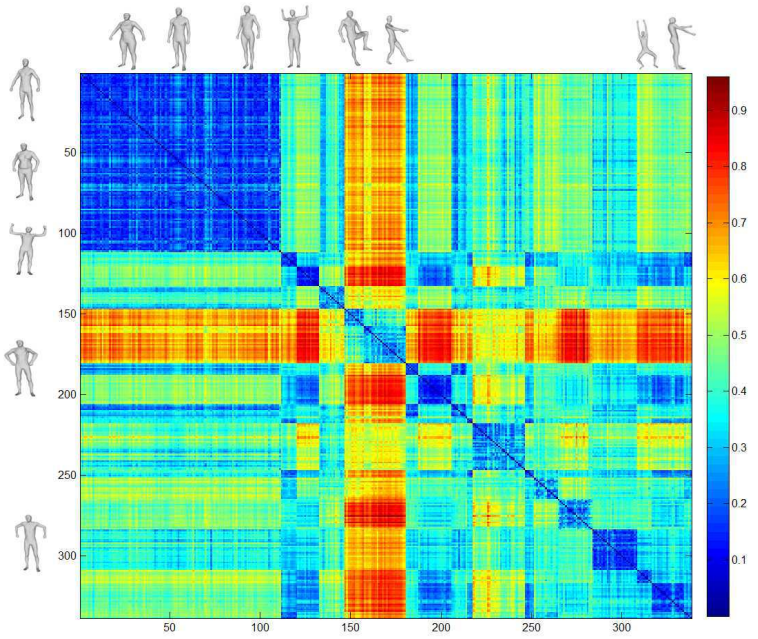


Figure 10: Confusion similarity matrix. The matrix contains pose dissimilarity computation between models of a 3D humans in different poses. More the color is cold more the two poses are similar.

Approach	NN(%)	FT(%)	ST(%)	E-Measure(%)
SH	71.0	57.9	75.5	41.3
EHC 10 curves	80.3	75.5	85.2	42.5
EHC 5 curves	**84.8**	**77.2**	**89.1**	**43.0**

Table 2: Retrieval statistics for pose based retrieval experiment

twist their torso to look around.

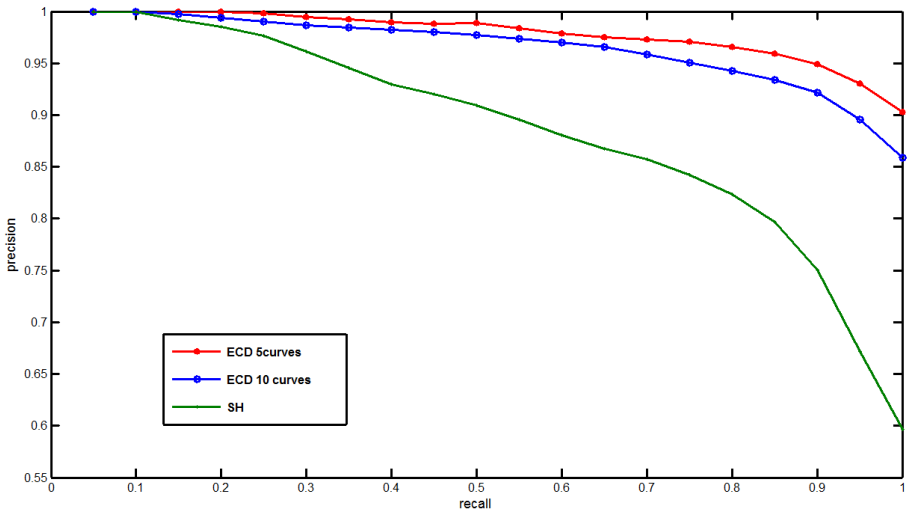


Figure 11: Precision-recall plot for pose-based retrieval.

7.3. Temporal shape similarity for 3D video sequences

We firstly present the protocol and the dataset used in these experiments and then, the results following this protocol are analyzed and compared to the most relevant state-of-the-art approaches.

7.3.1. Evaluation methodology

The recognition performance of the temporal shape descriptor is evaluated using a ground-truth dataset from a synthetic 3D video sequences proposed by Huang et al. [12] and a real captured 3D video sequences of people [49]. As described in Table 1 (2^{nd} raw), the synthetic data is obtained by 14 people (10 men and 4 women) performing 28 motions. Each sequence is composed of 100 frames and the whole dataset contains a total of 39200 frames.

Given the known correspondences, a temporal ground-truth similarity is computed between each two surfaces. The known correspondence is only

used to compute this ground truth similarity. Having two Mesh X and Y with N vertices $x_i \in X$ and $y_i \in Y$, a temporal-ground truth C_T is computed by combining a shape similarity C_p and a temporal similarity C_v as follows:

$$\begin{aligned} C_T(X,Y) &= (1-\alpha)C_p(x_i,y_j) + \alpha C_v(x_i,y_j) \\ C_p(X,Y) &= \tfrac{1}{N}\textstyle\sum_{k=1}^{N} d(x_i,y_j) \\ C_v(X,Y) &= \tfrac{1}{N}\textstyle\sum_{k=1}^{N} d(\dot{x}_i,\dot{y}_j) \end{aligned} \tag{8}$$

where d is an Euclidean distance, $\dot{x}_i$ and $\dot{y}_j$ are the derivation of x and y between next and current frame. the parameter α is used to balance the equation and it is set to 0.5 . In order to identify frames as similar or dissimilar, the temporal ground truth similarity matrix is binarized using a threshold set to 0.3 similarly to Huang et al. [12].

Finally, recognition performance is evaluated using the Receiver-Operator-Characteristic (ROC) curves, created by plotting the fraction of true-positive rate (TPR) against the fraction of false-positive rate (FPR), at various threshold settings. The true and false dissimilarity compare the predicted similarity between two frames, against the ground-truth similarity.

An example of self-similarity matrix computed using temporal ground-truth descriptor, static and temporal descriptors are shown in Figure 12. This figure illustrates also the effect of time filtering with increasing temporal window size for EHC descriptors on a periodic walking motion.

7.3.2. Result analysis

A comparison is made between our Temporal Extremal Human Curve (TEHC) and several descriptors from the state-of-the-art: Shape Distribution (SD) , Spin Image (SI) , Spherical Harmonics Representation (SHR),

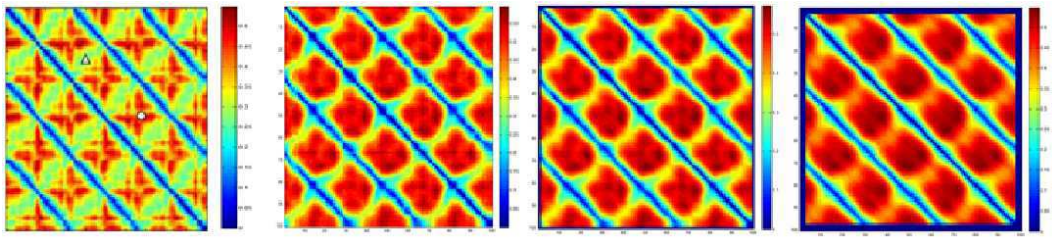


Figure 12: Similarity measure for "Fast Walk" motion in a straight line compared with itself. Coldest colors indicate most similar frames. 1^{st} matrix: temporal Ground-Truth (TGT). 2^{nd}, 3^{rd} and 4^{th} matrix: self-similarity matrix computed with Temporal EHC with window size 3, 5 and 7 respectively.

two Shape-flow descriptors, the global / local frame alignment Shape Histograms (SHvrG / SHvrS) (Huang et al. [12]) and Reeb-Graph as skeleton based shape descriptors (aMRG) (Tung et al. [51]). Note that a spectral representation was also evaluated by Huang et al. [21] which is the Multi-Dimentional Scaling (MDS). Huang et al. [12] evaluated the performances of all these descriptors for the purpose of shape similarity.

The effectiveness of our descriptor have been evaluated by varying temporal window and comparing it to the most relevant state-of-the-art descriptors [12] as shown in the plot of ROC curves in Figure 13.

Several observations can be made on the obtained results: (i) Our descriptor outperforms classic shape descriptors (SI, SHR, SD) and shows competitive results with SHvrS and aMRG. We also notice that recognition performance of EHC increases with the increase of the window size of time-filter like any other descriptor. In fact, time-filter reduces the minima in the anti-diagonal direction, resulting from motion in the static descriptor (Figure 13).

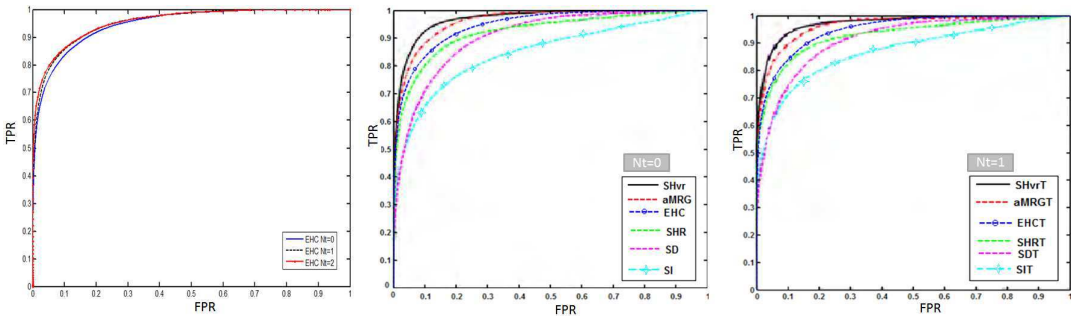


Figure 13: Evaluation of ROC curve for static and time-filtered descriptors on self-similarity across 14 people doing 28 motions. From right to left: ROC curves obtained by our TEHC descriptor with three different values of windows size Nt, ROC curve obtained by our EHC descriptor compared to different algorithms and ROC curves obtained with Nt= 1.

Multiframe shape-flow matching required in SHvrS allows the descriptor to be more robust but the computational cost will increase by the size of selected time window.

(ii) EHC descriptor by its simple representation, demonstrates a comparable recognition performance to aMRG. It is efficient as the curve extraction is instantaneous and robust as the curve representation is invariant to elastic and geometric changes thanks to the use of the elastic metric.

(iii) The result analysis for each motion shows that EHC gives a smooth rates that are stable and not affected by the complexity of the motion. Such complex motions are rockn'roll, vogue dance, faint, shot arm (Figure 14). However, this is not the case for SHvrS where performance recognition falls suddenly with complex motions as illustrated in Figure 15.

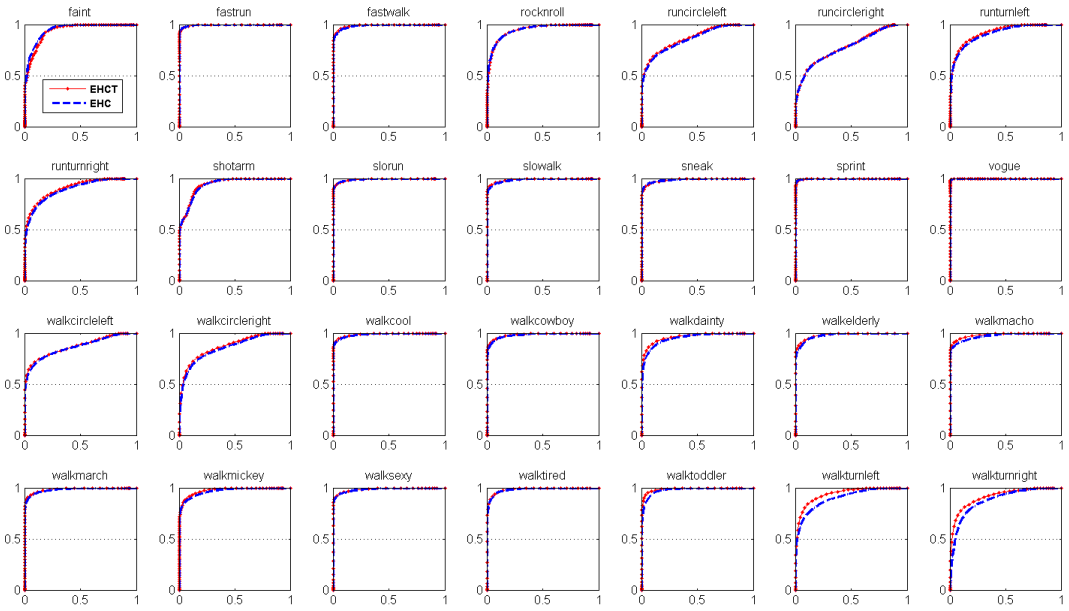
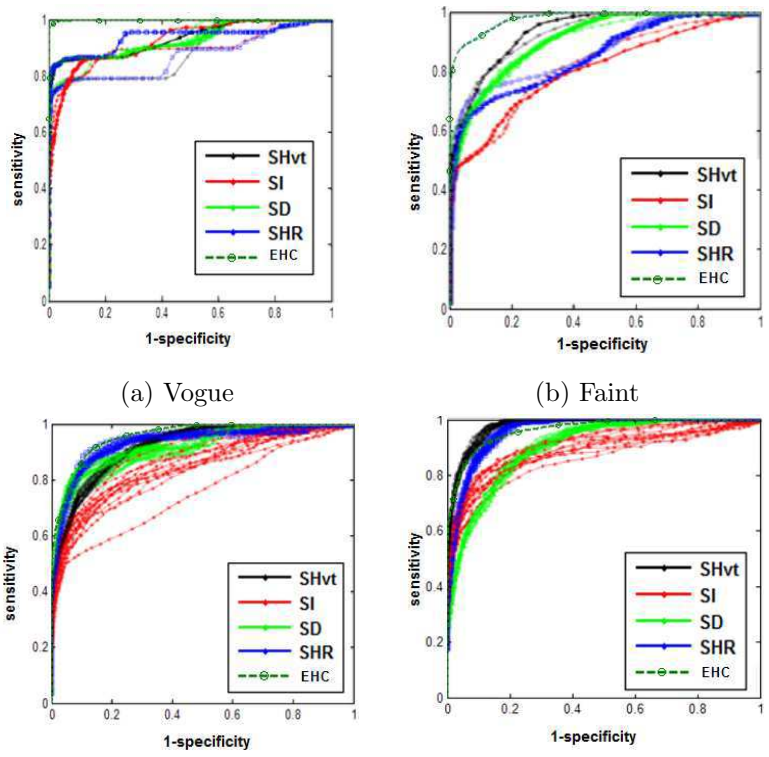


Figure 14: Evaluation of EHC descriptor against Temporal Ground Truth (TGT) for Nt=0 and Nt=1. ROC performance for 28 motions across 14 people.

We also applied the time filtering EHC descriptor on two real captured 3D video sequences of people. The first sequence is extracted from the dataset [49] described in Table 1 (3^{rd} row). The second one is extracted from real data reconstructed by multiple camera video [50] and described in Table 1 (4^{th} row).

Inter-person similarity across two people in a walking motion with an example similarity curve are shown in Figure 16 (a). Our temporal similarity measure identifies correctly similar frames across different people. These similar frames are located in the minima of the similarity curve. In addition,



(a) Vogue (b) Faint

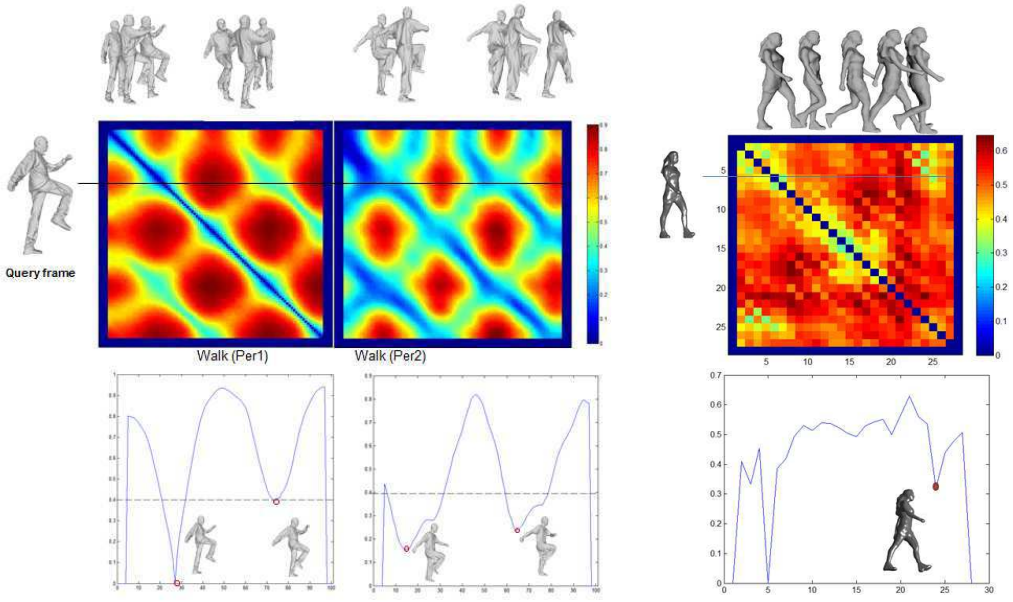
(c) Shot arm (d) Rockn'roll

Figure 15: Evaluation of ROC curves for complex motions with Nt=3.

despite the topology change and the reconstruction noise as shown in Figure 16 (b), our algorithm succeed to identify correctly the frame in the sequence similar to the query.

7.4. Motion segmentation and retrieval

In this section, we evaluate temporal shape similarity descriptor. Details about the computation of the ground truth descriptor are given in addition to the description of the different datasets used for evaluation. The results obtained by our approach, compared to those of different state-of-the-art



(a) (b)

Figure 16: Inter-person similarity measure for real sequences. Similarity matrix, curve and example frames for (a) Walk motion across two actors [49] (b)walk motion for Roxanne [50] Game Character Walk .

descriptors, are then discussed.

7.4.1. Evaluation methodology

The two datsets (2) and (3) presented in Table 1 are used in these experiments. From the synthetic dataset [12], we have chosen 14 different motions: walk (slow, fast, circle left/right, cowboy, march, mickey), run (slow, fast, circle left/right), sprint, and rockn'roll. These motions are performed by two actors (a woman and a man) making a total of 28 motions (2800 frames). They are chosen for their interesting challenges as: (i) change in execution

rate (slow/fast motions) (ii) change in direction while moving (walking in straight line, moving in circle and turning left and right) (iii) change in shape (a woman and a man). We used these motion sequences for both segmentation and retrieval experiment.

To validate the segmentation step, we segment all these 3D video sequences with the proposed approach and then compare results to manual segmentation ground-truth. In the retrieval process, each query clip is compared to all other clips obtained by the segmentation of sequences. Finally, the statistics (NN, FT, ST and E-measure) are used for the evaluation.

7.4.2. Analysis of motion segmentation result

Plotting the distance between EHC representation of successive frames gives a very noisy curve. The break points from this curve do not define semantic clips and the extracting of minima leads to an over-segmentation of the sequence (see Figure 17 (top)). To obtain more significant local minima, we convolve the curve with a time-filter allowing to take into account the motion variation, not only between two successive frames but also in a time window. The motion degree after convolution is shown in Figure 17 (bottom). Break points are more precise and delimits significant clips corresponding to step change in the video sequence. In order to evaluate its efficiency, we apply our segmentation method on the whole dataset (3) described in Table 1 (3^{rd} raw) and then compare the results to a manual segmentation of the base done carefully .

We performed the clip segmentation for all window size values from 1 to 11 over a representative set of clips extracted from the dataset (3) [49]. Compared to manual ground truth, the best segmentation is obtained using

a window size of 5. This value is then fixed for the rest of the tests. The segmentation of the dataset(3) gives 83 segmented clips (78 correct clips and 5 incorrect clips). This can be explained by the fact that the 5 failing clips are short. They contain about 6 frames at most and do not describe atomic significant actions. Otherwise, the a total of 144 clips have been obtained by the segmentation of the 14 motions taken from the dataset (2) described in Table 1 (2^{nd} raw) performed by two actors.

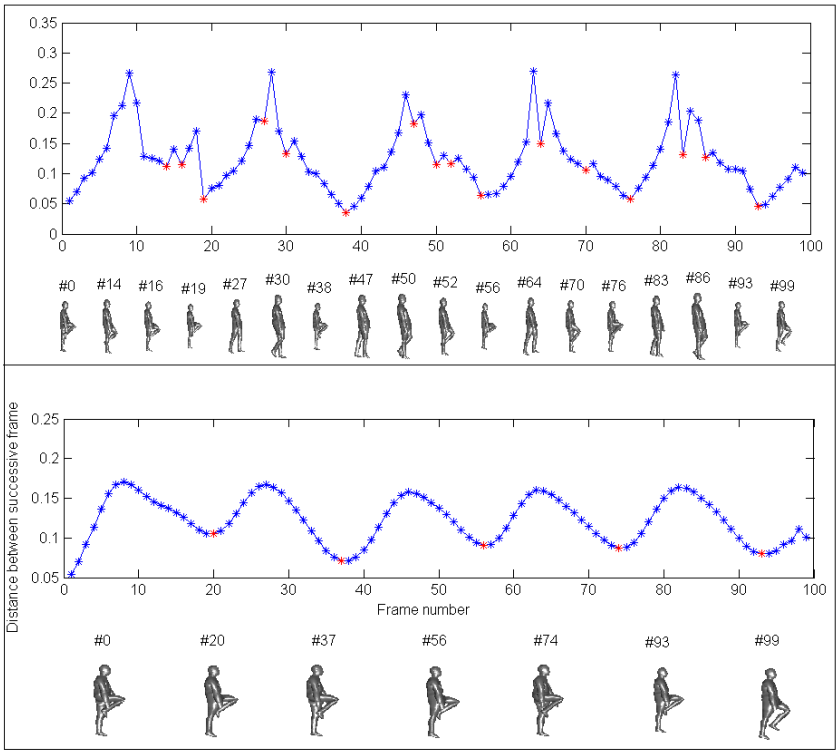


Figure 17: Speed curve smoothing.

Figure 18 shows some results of motion segmentation on a "slow walk" and a "fast walk" motions. Although the walk speed increases, the motion

segmentation remains significant and does not change and corresponds to the step change of the actor. The Rockn'roll dance motion segmentation is also illustrated in Figure 18 (bottom). Thanks to the selection of local minima in a precise neighborhood, only significant break points are detected.

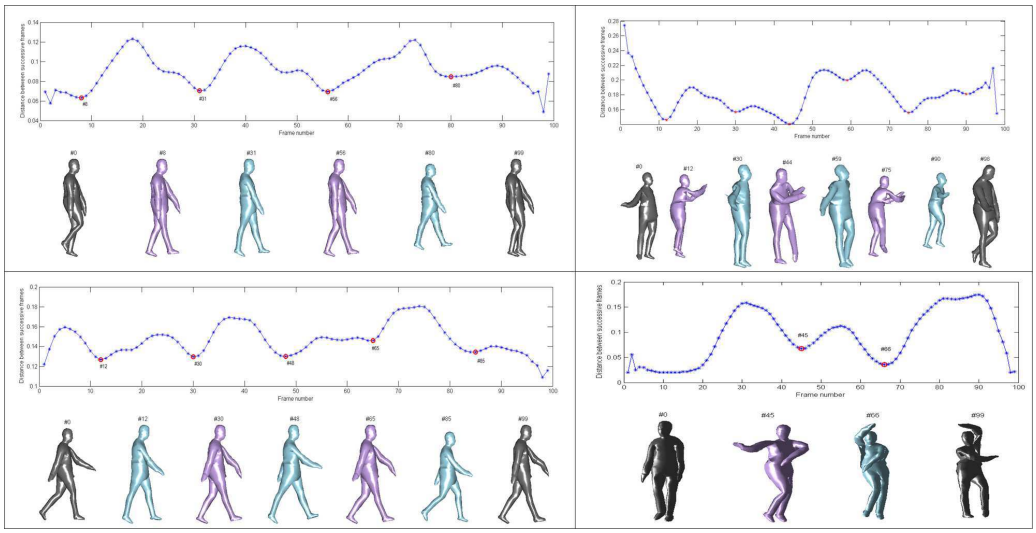


Figure 18: Various motion segmentation. From right top to left bottom, motions are: slow walk, Rockn'roll dance, fast walk, vogue dance.

7.4.3. Analysis of motion retrieval result

The motion segmentation method, applied on 14 motion sequences from the dataset (2) and performed by a man and a woman, gives a total of 144 clips. These clips, with an average number of frames per clip equal to 15, are categorized into 14 classes. The motion sequences consist mainly of different styles of walking, running and some dancing sequences. Classes grouped together represent different styles of walking, running and dancing steps. For example, a step change in a walk may represent a class and groups

similar clips done with different speed and in different trajectories. We notice that right to left change step is grouped in a different class than left to light change step.

The similarity metric represented by elastic measure values between each pair of clips allows us to generate a confusion matrix for all classes of clips, in order to evaluate the recognition performance by computing dynamic retrieval measures thanks to a manually annotated ground truth. An example of the matrix representing the similarity evaluation score among clips in sequences performed by a female actress against the clips of sequences of motions performed by a male actor is shown in Figure 19. More the color is cold more the clips are similar.

Thanks to the use of DTW, it is noticed that similarity score between same clips done in different speeds is small (see Figure 19). The matching between the clip representing change in step in a slow walk motion composed of 25 frames and a fast walk motion, composed of 18 frames, is small.

Besides, our approach succeed to retrieve clips within motions done in different ways. For example, the walk circle clips can be matched with the clips of slow walk motion done in a straight line (see Figure 19). This explains why the use of an elastic metric, to compare and match trajectories, makes the process independent to rotation. Although the actors performing the motions are different, it is observed that similar clips yield smaller similarity score. Like it is shown in "Rockn'roll" dance motion, steps of the dance performed by different actors are correctly retrieved.

It is demonstrated that 79.26% of similar motion clips are included in the first tier and 93% of clips are correctly retrieved in the second tier. It

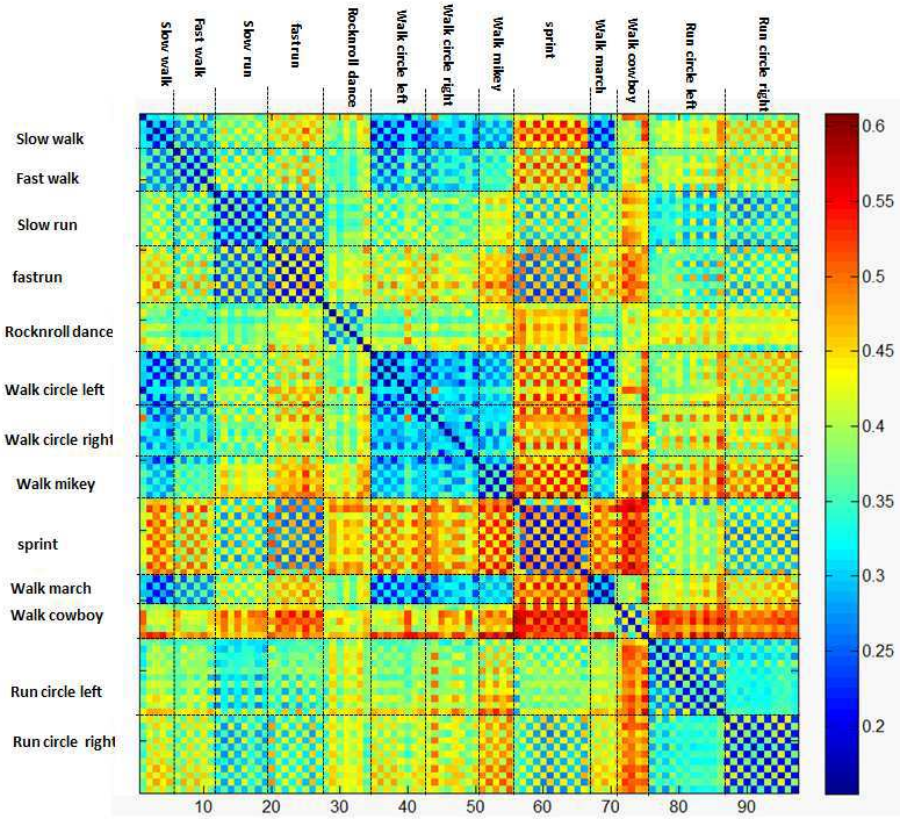


Figure 19: Similarity matrix evaluation between clips. More the color is cold more the two clips are similar.

is a rather good performance considering that only such low-level feature as the EHC is utilized in the matching. This can be explained by the fact that geodesics are not completely invariant to the topology changes. Thereby, the extracted sequential curves that represent the trajectory tend to change the path on the models for certain motions and therefore mislead the matching performed by DTW.

We also apply our retrieval approach to a real captured 3D video sequence

from the real dataset (3) described in Table 1 (3^{rd} raw). Self similarity example with an actor in a walking motion (walking in circular way) and its similarity curve are shown in Figure 20. For the query clip presented at the right of the figure, retrieved clips are found correctly in the sequence when the actor is turning.

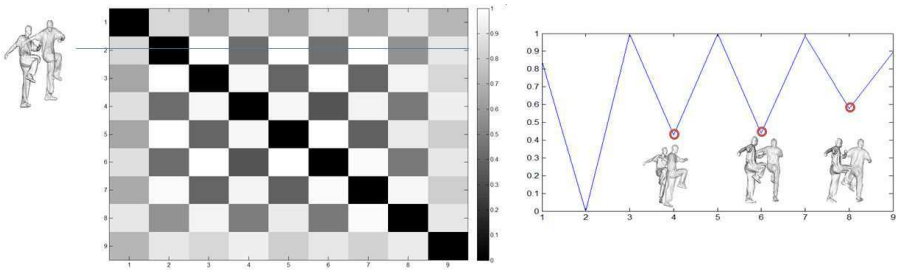


Figure 20: Experimental results for 3D video retrieval using motion of "walk in circle".

7.5. Data summarization and content-based retrieval

In this section, we firstly conducted multiple experimental trials by analyzing the video clustering method on two aspects: the pose-based clustering and the clip-based clustering. Secondly, we evaluate the impact of the summarization process on the retrieval system by comparing the results with and without using clustering.

7.5.1. Content-based summarization

The performance of the content-based summarization approach is evaluated for pose and clip data. To validate the pose-based summarization, we use a composed long sequence of a subject performing walk and squat

motions from the dataset (3). For clip-based summarization experiment, the same 28 motions used for video segmentation and the retrieval have been used.

The effectiveness of clustering process is evaluated by the number of clusters found which should allow the identification of eventual redundant patterns. The threshold Th in the Algorithm 2 is set accordingly to the values of the similarity function. The distances computed between descriptors (EHC for pose and trajectory of EHC for clip) are normalized to return values in the range [0 1], and Th was then defined experimentally. An optimal setting of Th should return a set of clusters similar to what a "hand-made" ground-truth classification would perform. The Figure 21 shows the clustering result obtained from the composed long sequence. The number of clusters decreases with the increase of the threshold Th. We obtain the best result for $Th = 0.5$ with 51 clusters partitioned as the bar diagram shown in the right of the Figure 21.

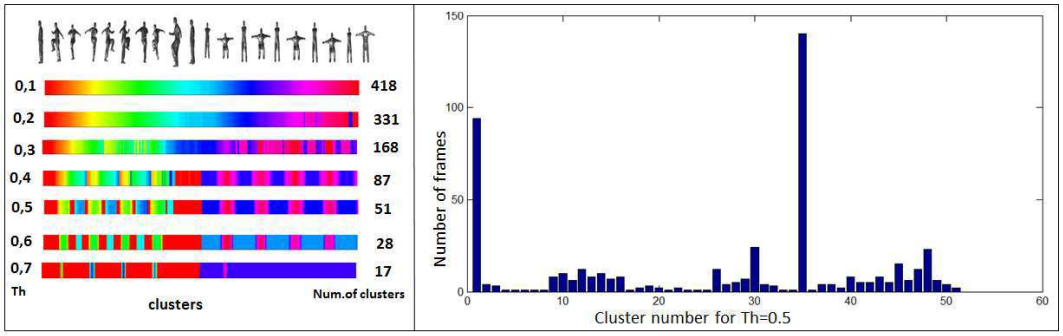


Figure 21: Frame clustering process with respect to threshold Th.

Pose-based clustering process can be improved by increasing the window

size of the time filter as shown in Figure 22.

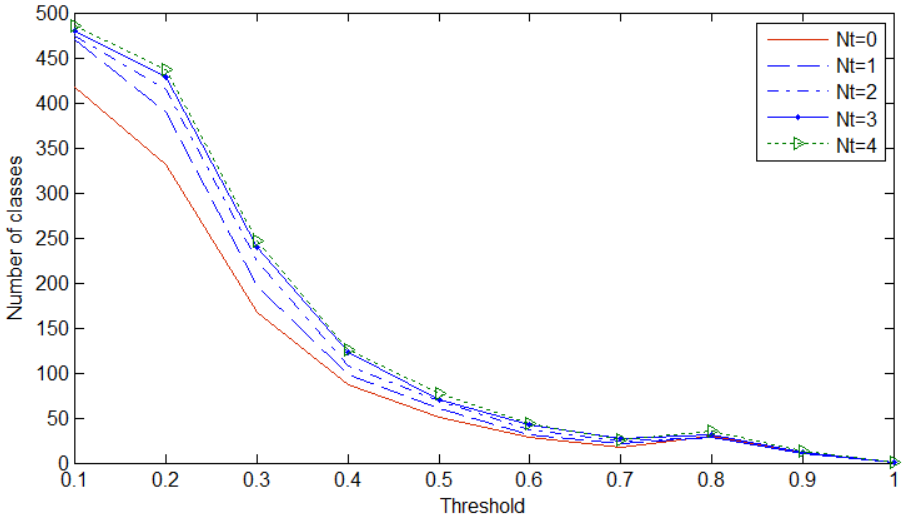


Figure 22: Frame clustering with respect to a threshold and with different window size.

We notice from this figure that for a $Th = 0.2$, the number of clusters varies from 330 to 440 and a good compromise is obtained for $Nt = 3$.

Furthermore, clustering is applied on 14 motions extracted from the dataset (3) and performed by two actors (a man and a woman) in order to evaluate the efficiency of the clip-based clustering. By decreasing the threshold Th of the clustering algorithm, we obtain more clusters. Experimentally, we set Th to 0.43 and obtain 23 clusters from initially 110 clips for the first actor and 26 clusters for the second one (see Figure 23). We notice that clips representing sprint or running steps are clustered together.

The video summarization process can be used efficiently in hierarchical structure, starting by video segmentation into clips, followed by clip-based

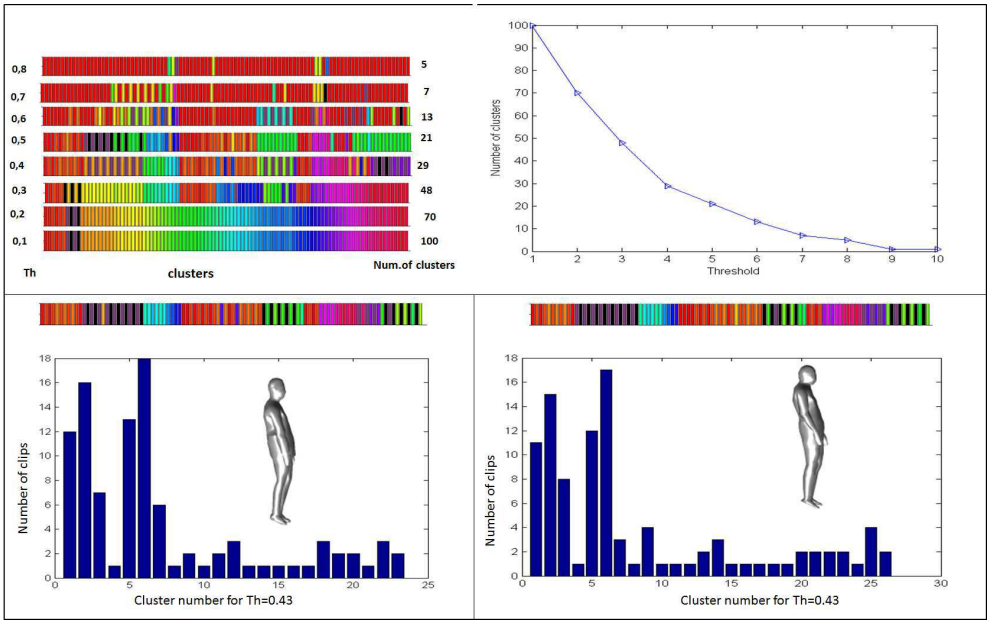


Figure 23: Clustering clips from a sequence of two actors performing 14 motions from the dataset (3) for a total of 1400 frames, with respect to Th. In second raw, the variation of clip number in each cluster is presented.

clustering and then a pose-based clustering performed on the frames of all represented clusters of the clips resulting from the last step. The effectiveness of our summarization process is shown in Figure 24 for the sequence of a real actor performing walking and squatting motion. From 500 frames segmented into 18 clips, the clustering process gives 6 clusters. The new subsequence containing 6 clips (most representative clip in each cluster) and 180 frames is then clustered into 41 clusters where each one represent a class of pose.

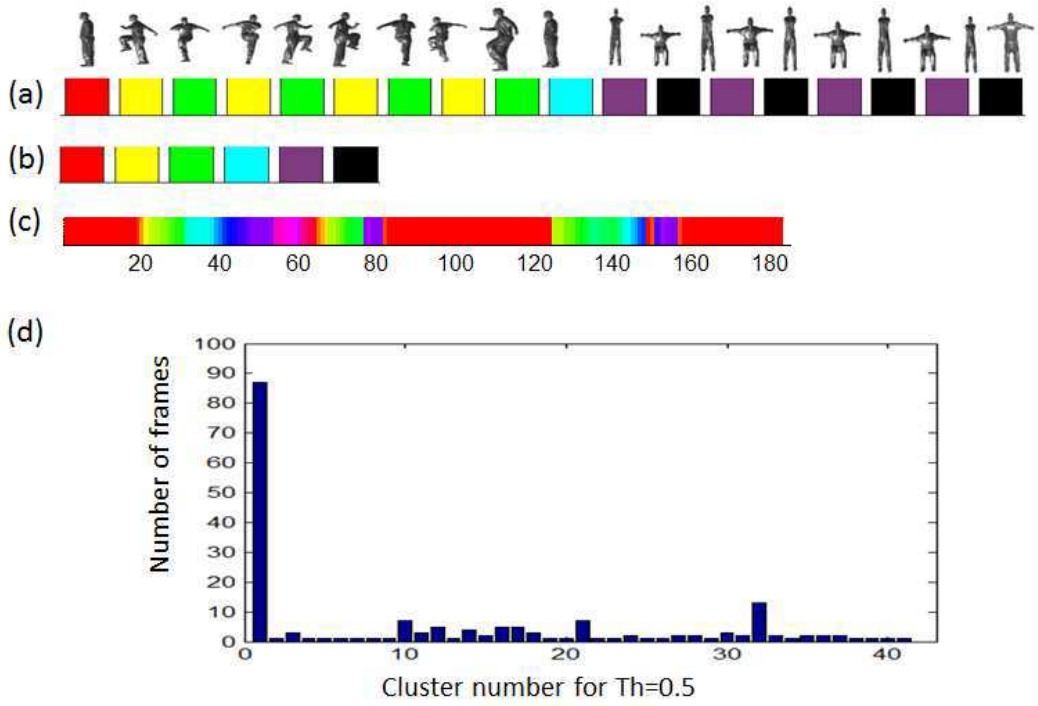


Figure 24: Summarization process: (a) for a sequence of 500 frames segmented into 18 clips, the clustering process returns 6 clusters of clips using $Th = 0.38$ (b) subsequence of clustered clips (180 frames) where each cluster is represented by only one clip chosen as the Karcher Mean clip of the cluster, (c) clustering of subsequence into 41 clusters of frames using $Th = 0.5$, (d) distribution of the number of frames in clusters.

7.5.2. Hierarchical data retrieval

For a mesh model of 1 MB size, the size of the 3D video sequence grows linearly of 1 MB per frame. Hence, the video retrieval becomes very difficult in long sequences. Within our framework, we propose to combine the data clustering approach with the content-based retrieval in order to perform an

hierarchical retrieval.

The clustering approach gathers models with similar poses/clips in clusters. If we consider the element of cluster as a pose, clusters are firstly performed over the entire sequence in order to gather frames with similar poses and then a template model is obtained for each cluster by computing its Karcher Mean as described in section 4.3. The retrieval system can then be described as an hierarchical structure composed of two levels, the first one containing templates and the second one containing all models of the dataset. In view of this structure, a natural way is to start at the top, compare the query with the template of each cluster and proceed down the branch that leads to the closest shape.

We reconsider the same experiments for pose based retrieval in section 7.2 by applying the hierarchical approach to the dataset summarized in Table 1 (1^{st} raw) . Each query model is compared to each one of the template models representing the clusters. The elastic measure values are used to generate a confusion matrix for all classes of pose, in order to evaluate the recognition performance by computing statistic retrieval measures thanks to the provided ground truth. The matrix of comparison in the first level (model-template comparison), is shown in Figure 25.

If we compare this matrix to that already obtained for the same dataset without the use of summarization (Figure 10), you can easily notice the effectiveness of the summarization. The main advantage of this approach is the reduction of computation time which complexity pass from n to $log(n)$ while keeping relevant information. Retrieval performances obtained from this matrix for FT, ST and E-Measure are respectively 84.5% , 88.2% and

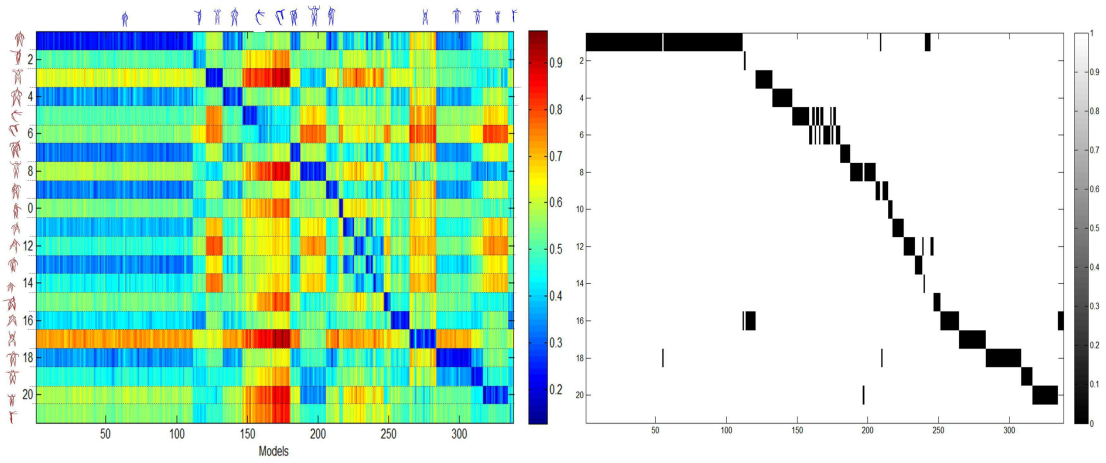


Figure 25: Similarity matrix and its binarization for template pose of each class against all models in the dataset.

43.6%. Comparing these results to those in Table 2, a small improvement is achieved for classic retrieval scenario in term of second tier.

In term of pose classification, the obtained accuracy is about 90.24%. Models of the class ♯2 are the most ones affected by misclassification and are assigned to the class ♯16. Looking at these two classes, we perceive that their poses are close to each other, both represent people with hands outstretched. The only difference is that one does with open legs and the other with closed ones.

Finally, if we consider the element of cluster as a clip, where a video segmentation is firstly performed on the whole sequence. In this case, the template model is a "mean" clip obtained for each cluster of clips by computing its Karcher Mean (see Algorithm 1). The retrieval system can then

be viewed as above with hierarchical structure. As experimental test, we performed a similar experimentation on the 14 motions performed by two actors as already evaluated in the section 7.3. In this experimentation, each query is a clip compared to each one of the template models representing the clusters of clips. The similarity measure values obtained by DTW algorithm between clips are used to generate a confusion matrix for all classes of clips, in order to evaluate the recognition performance by computing statistic retrieval measures thanks to the provided ground truth. The matrix of comparison in the first level (model-template comparison) is shown in Figure 26.

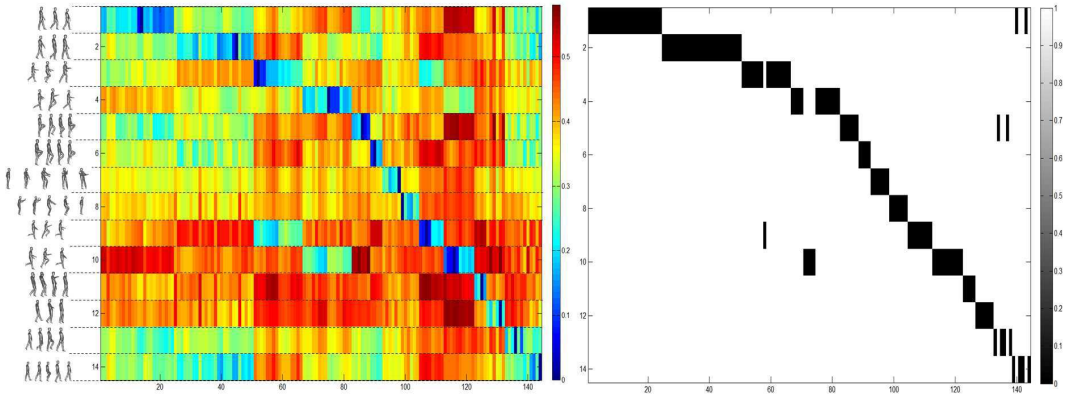


Figure 26: Similarity matrix and its binarization for template clip of each class against all clips in the dataset.

Retrieval performances obtained from this matrix for FT, ST and E-Measure are respectively 84.09%, 95.83% and 55.26%. In term of clip classification, obtained accuracy is about 93.75%. The analysis of the result given by the binarized matrix shows that the most misclassified clips are those of

"fast run" class. In fact, they are assigned to class template representing "sprint" motion class.

8. Discussion

The advantages of using EHC to represent human pose and motion in our approach include: (1) invariance to affine transformation (2) possibility to compute mean poses and mean clips (3) the use of well defined measure for pose comparison in Reimannian manifold and (4) the use of well established algorithm of DTW to align sequences taking benefits from the temporal aspect of curves.

However, this representation has some limitations. Firstly, EHC depends on the accuracy of extremities (head and limbs) extraction and on the definition of the path connecting end-points. In fact, the extraction of end-points and extremal curves is based on the definition of geodesic distance between each pair of curves. Thus, geodesic distances play an important role in our geometric representation of the human body shape. However, they are sensitive to significant topology changes as shown in Figure 27. In this figure, only 4 extremities are successfully detected and the left hand extremity is missed. Thus, information about position of this hand is lost. In the future, other strategies will be investigated for the extremities extraction step and shortest path detection on the mesh by using diffusion or commute time distances as presented by Elkhoury et al. [15] and Sun et al. [52].

Secondly, we note that our curve extraction can be sensitive to loose clothes. For example, the mesh represented in Figure 27 shows a girl wearing a skirt and the shape of the curve connecting her feet is different from the

same curve extracted on her mesh if she were wearing a trouser. This problem will be even more critical if she wears a long skirt.

Thirdly, a prior knowledge on the direction of the posture of the human body for the starting frame in video sequence is used to distinguish between left/right hand and foot. Other feature matching algorithms, like Heat Kernel Signature as proposed by Sun et al. [52] and Zheng et al. [53], could be used in future work to correctly identify the right from the left side.

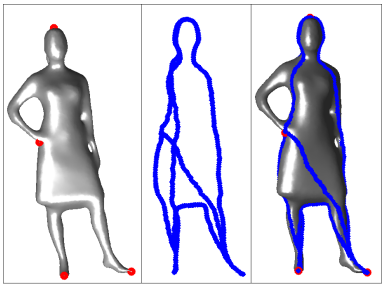


Figure 27: Example of failed extraction of EHC in presence of a topological change.

9. Conclusion

In this work, we have proposed a unified framework able to represent human body shape with a pose descriptor, as well as a sequence of frames with a specific representation. This framework relies on an Extremal Human Curve descriptor (EHC), based on extremal features and geodesics between each pair of them. This descriptor has the advantage of being a skeletal representation, which is trackable over time. It is also an extremal descriptor of the surface deformation which is composed by a collection of local open-

3D-curves. The representation of these curves and the comparison between them are performed in the Riemannian shape space of open curves. By this way, we have chosen to represent the pose of a mesh regardless to its rotation, translation and scale. Convoluted with a time filter to incorporate the motion, it becomes a temporal descriptor for pose retrieval. The degree of motion using feature vector, extracted from this descriptor, is then used for spliting continuous sequence into elementary motion segments called clips. Each clip describing an atomic movement is characterized by EHC representation associated to human mesh. The open curves in 3D space, which are the elements of EHC representation, are viewed as a point in the shape space of open curves and hence each clip is represented by a trajectory on this space. Dynamic time warping is used to align different trajectories and give a similarity score between each two clips.

The quality of our descriptor regarding the recognition performance of shape similarity in 3D video is analyzed and verified also by comparison with other related recent techniques. Moreover, our approach achieves a performance accuracy of 93.44% for video retrieval as second tier, which is encouraging. Finally, we will investigate 3D human action recognition and semantic activity analysis based on this framework.

10. Acknowledgements

This research program is supported partially by the Region Nord-Pas de Calais (France).

References

[1] K. M. Cheung, S. Baker, T. Kanade, Shape-from-silhouette across time part i: Theory and algorithms, in: International Journal of Computer Vision, Vol. 62, 2005, pp. 221 – 247.

[2] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, S. Thrun, Performance capture from sparse multi-view video, in: ACM SIGGRAPH, Vol. 27, ACM, 2008, pp. 1–10.

[3] T. Kanade, P. Rander, P. J. Narayanan, Virtualized reality: Constructing virtual worlds from real scenes, in: IEEE MultiMedia, Vol. 4, 1997, pp. 34–47.

[4] T. Tung, S. Nobuhara, T. Matsuyama, Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo, in: IEEE 12th International Conference on Computer Vision (ICCV), 2009, pp. 1709 –1716.

[5] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, R. Szeliski, A comparison and evaluation of multi-view stereo reconstruction algorithms, in: IEEE International Conference on Computer Vision and Pattern Recognition, Vol. 1 of CVPR '06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 519–528.

[6] A. E. Johnson, M. Hebert, Using spin images for efficient object recognition in cluttered 3d scenes, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, pp. 433–449.

[7] R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin, Shape distributions, in: ACM Transactions on Graphics, Vol. 21, 2002, pp. 807–832.

[8] M. Ankerst, G. Kastenmüller, H.-P. Kriegel, T. Seidl, 3d shape histograms for similarity search and classification in spatial databases, in: Proceedings of the 6th International Symposium on Advances in Spatial Databases, SSD '99, 1999, pp. 207–226.

[9] P. Huang, A. Hilton, Shape-colour histograms for matching 3d video sequences, in: IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2009, pp. 1510 –1517.

[10] M. Körtgen, G. J. Park, M. Novotni, R. Klein, 3d shape matching with 3d shape contexts, in: The 7th Central European Seminar on Computer Graphics, 2003.

[11] M. Kazhdan, T. Funkhouser, S. Rusinkiewicz, Rotation invariant spherical harmonic representation of 3d shape descriptors, in: Proceedings of the Eurographics/ACM SIGGRAPH symposium on Geometry processing, SGP '03, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 2003, pp. 156–164.

[12] P. Huang, A. Hilton, J. Starck, Shape similarity for 3d video sequences of people, in: International Journal of Computer Vision, Vol. 89, Kluwer Academic Publishers, Hingham, MA, USA, 2010, pp. 362–381.

[13] T. Tung, T. Matsuyama, Topology dictionary for 3d video understanding, in: IEEE Transactions on Pattern Analysis and Machine Intelli-

gence, Vol. 34, IEEE Computer Society, Los Alamitos, CA, USA, 2012, pp. 1645–1657.

[14] H. Tabia, M. Daoudi, J.-P. Vandeborre, O. Colot, A new 3d-matching method of nonrigid and partially similar models using curve analysis, in: IEEE Trans. Pattern Anal. Mach. Intell., Vol. 33, 2011, pp. 852–858.

[15] R. El Khoury, J.-P. Vandeborre, M. Daoudi, Indexed heat curves for 3d-model retrieval, in: 21st International Conference on Pattern Recognition (ICPR 2012), Tsukuba Science City, Japon, 2012, pp. 1964–1967.

[16] S. Mahmoudi, M. Daoudi, A probabilistic approach for 3d shape retrieval by characteristic views, in: Pattern Recognition Letters, Vol. 28, 2007, pp. 1705 – 1718.

[17] J. Sun, M. Ovsjanikov, L. Guibas, A concise and provably informative multi-scale signature based on heat diffusion, in: Proceedings of the Symposium on Geometry Processing, SGP '09, 2009, pp. 1383–1392.

[18] Y. Lipman, T. Funkhouser, Mobius voting for surface correspondence, in: ACM SIGGRAPH 2009 Papers, Vol. 28 of SIGGRAPH '09, 2009, pp. 1–12.

[19] M. Ovsjanikov, Q. Mérigot, F. Mémoli, L. J. Guibas, One point isometric matching with the heat kernel, in: Comput. Graph. Forum, Vol. 29, 2010, pp. 1555–1564.

[20] J. Lee, J. Chai, P. S. A. Reitsma, J. K. Hodgins, N. S. Pollard, Interactive control of avatars animated with human motion data, in: Proceed-

ings of the 29th annual conference on Computer graphics and interactive techniques, ACM, New York, NY, USA, 2002, pp. 491–500.

[21] P. Huang, T. Tung, H. Nobuhara, S.and Hilton, T. Matsuyama, Comparison of skeleton and non-skeleton shape descriptors for 3d video, in: Proceedings of the Fifth International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT'10), Pairs, France, 2010.

[22] T. Yamasaki, K. Aizawa, Motion segmentation and retrieval for 3d video based on modified shape distribution, in: Journal on Applied Signal Processing EURASIP, Vol. 2007, Hindawi Publishing Corp., New York, NY, United States, 2007, pp. 211–211.

[23] J. Wang, H. Zheng, View-robust action recognition based on temporal self-similarities and dynamic time warping, in: Computer Science and Automation Engineering (CSAE), 2012 IEEE International Conference on, Vol. 2, 2012, pp. 498–502.

[24] L. Wang, L. Cheng, L. Wang, Elastic sequence correlation for human action analysis, in: Image Processing, IEEE Transactions on, Vol. 20, 2011, pp. 1725–1738.

[25] Y.-S. Tak, J. Kim, E. Hwang, Hierarchical querying scheme of human motions for smart home environment, in: Engineering Applications of Artificial Intelligence, Vol. 25, Pergamon Press, Inc., Tarrytown, NY, USA, 2012, pp. 1301–1312.

[26] M. Mortara, G. Patane, Affine-invariant skeleton of 3d shapes, in: Pro-

ceedings of the Shape Modeling International (SMI'02), SMI '02, IEEE Computer Society, Washington, DC, USA, 2002, pp. 245 – 252.

[27] S. Katz, G. Leifman, A. Tal, Mesh segmentation using feature point and core extraction, in: The Visual Computer, Vol. 21, Springer-Verlag, 2005, pp. 649–658.

[28] J. Tierny, J.-P. Vandeborre, M. Daoudi, Invariant high level reeb graphs of 3d polygonal meshes, in: International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT), IEEE Computer Society, Los Alamitos, CA, USA, 2006, pp. 105–112.

[29] F. Lazarus, A. Verroust, Level set diagrams of polyhedral objects, in: Proceedings of the fifth ACM symposium on Solid modeling and applications, SMA '99, New York, NY, USA, 1999, pp. 130–140.

[30] M. F. Abdelkader, W. Abd-Almageed, A. Srivastava, R. Chellappa, Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds, in: Journal of Computer Vision and Image Understanding, Vol. 115, Elsevier Science Inc., New York, NY, USA, 2011, pp. 439–455.

[31] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, R. Slama, 3d face recognition under expressions, occlusions, and pose variations, in: IEEE Trans. Pattern Anal. Mach. Intell., Vol. 35, 2013, pp. 2270–2283.

[32] S. Joshi, E. Klassen, A. Srivastava, I. Jermyn, A novel representation for riemannian analysis of elastic curves in rn, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07), 2007, pp. 1 –7.

[33] P. W. Michor, D. Mumford, Riemannian geometries on spaces of plane curves, in: J. Eur. Math. Soc., Vol. 8, 2006, pp. 1–48.

[34] E. Klassen, A. Srivastava, W. Mio, S. Joshi, Analysis of planar shapes using geodesic paths on shape spaces, in: IEEE Pattern Anal. Mach. Intell., Vol. 26, 2004, pp. 372–383.

[35] A. Yezzi, A. Mennucci, Conformal metrics and true "gradient flows" for curves, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, ICCV, IEEE Computer Society, Washington, DC, USA, 2005, pp. 913–919.

[36] A. Srivastava, E. Klassen, S. Joshi, I. Jermyn, Shape analysis of elastic curves in euclidean spaces, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, 2011, pp. 1415–1428.

[37] I. Koprinska, S. Carrato, Temporal video segmentation: A survey, in: Signal Processing: Image Communication, 2001, pp. 477–500.

[38] Y. Rui, P. Anandan, Segmenting visual actions based on spatio-temporal motion patterns, in: Proceedings. IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2000, pp. 111 –118 vol.1.

[39] T.-S. Wang, H.-Y. Shum, Y.-Q. Xu, N.-N. Zheng, Unsupervised analysis of human gestures, in: Proceedings of the Second IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing, PCM '01, Springer-Verlag, London, UK, UK, 2001, pp. 174–181.

[40] T. Shiratori, A. Nakazawa, K. Ikeuchi, Rhythmic motion analysis using motion capture and musical information, in: Proceedings of IEEE

International Conference on Multisensor Fusion and Integration for Intelligent Systems, 2003, pp. 89 – 94.

[41] K. Kahol, P. Tripathi, S. Panchanathan, Automated gesture segmentation from dance sequences, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 883 – 888.

[42] J. Xu, T. Yamasaki, K. Aizawa, 3d video segmentation using point distance histograms, in: IEEE International Conference on Image Processing (ICIP), Vol. 1, 2005, pp. 701–704.

[43] T. Yamasaki, K. Aizawa, Motion segmentation of 3d video using modified shape distribution, in: IEEE International Conference on Multimedia and Expo, 2006, pp. 1909 –1912.

[44] P. Huang, A. Hilton, J. Starck, Automatic 3d video summarization: Key frame extraction from self-similarity, in: Fourth International Symposium on 3D Data Processing, Visualization and Transmission, 2008, pp. 1 – 8.

[45] T. Giorgino, Computing and visualizing dynamic time warping alignments in r: The dtw package, in: Journal of Statistical Softwar, Vol. 31, 2009, p. 1–24.

[46] R. Slama, H. Wannous, M. Daoudi, Extremal human curves: a new human body shape and pose descriptor, in: Accepted - 10th IEEE International Conference on Automatic Face and Gesture Recognition, Shanghai, China, 2013, pp. 1–6.

[47] B. Allen, B. Curless, Z. Popović, The space of human body shapes: reconstruction and parameterization from range scans, in: ACM SIGGRAPH, Vol. 22, ACM, New York, NY, USA, 2003, pp. 587–594.

[48] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, H.-P. Seidel, A statistical model of human pose and body shape, in: P. Dutr'e, M. Stamminger (Eds.), Computer Graphics Forum (Proc. Eurographics 2008), Vol. 2, Munich, GermMortara:SMI:2002:Skeletonany, 2009, pp. 337–346.

[49] D. Vlasic, I. Baran, W. Matusik, J. Popović, Articulated mesh animation from multi-view silhouettes, in: ACM Siggraph, ACM, New York, NY, USA, 2008, pp. 97:1–97:9.

[50] J. Starck, A. Hilton, Surface capture for performance-based animation, in: Computer Graphics and Applications, Vol. 27, 2007, pp. 21–31.

[51] T. Tung, F. Schmitt, The augmented multiresolution reeb graph approach for content-based retrieval of 3d shapes, in: International Journal of Shape Modeling, Vol. 11, 2005, pp. 91–120.

[52] J. Sun, M. Ovsjanikov, L. Guibas, A concise and provably informative multi-scale signature based on heat diffusion, in: Proceedings of the Symposium on Geometry Processing, SGP '09, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 2009, pp. 1383–1392.

[53] Y. Zheng, C.-L. Tai, E. Zhang, P. Xu, Pairwise harmonics for shape analysis, Vol. 19, IEEE Computer Society, Los Alamitos, CA, USA, 2013, pp. 1172–1184.